NOBLEJACKAL / NOMOS

NOMOS

Протокол GEO-аудита NOMOS

МЕТОД GEO-1000

Протокол измерения представления сущностей в генеративных системах на уровне мирового населения

Kaan MURAZ

0.9.0 · РУССКАЯ РЕДАКЦИОННАЯ ВЕРСИЯ

Протокол GEO-аудита NOMOS

Протокол измерения представления сущностей в генеративных системах на уровне мирового населения

Автор и ответственный основатель: Kaan MURAZ

Идентичность метода: NOMOS by NobleJackal

Версия: 0.9.0 — кандидат окончательного текста, русская редакционная версия

Язык: русский (ru-RU)

© 2026 Kaan MURAZ. Лицензия на публикацию этой версии-кандидата ещё не объявлена.

В этой версии NOMOS представляет собой открытый версионируемый проект стандарта, переданный для независимого тестирования. Это не принятый международный стандарт и не действующая программа сертификации.

Статус реального бенчмарка: BQ-0 — не выполнен. Цифры в главе 18 — это расчётная синтетическая демонстрация метода, а не результаты с реальными пользователями, реальными ИИ-продуктами или независимого полевого исследования.

Эта версия сама по себе не создаёт знак соответствия NOMOS, заявление «NOMOS 950+», сертификат, рейтинг или стороннее подтверждение.

КАК ЧИТАТЬ ЭТУ КНИГУ

Эта книга — не руководство по выполнению GEO. Это протокол аудита и проект стандарта, который определяет условия популяции, доказательств, контекста, времени, вынесения решений и управления, при которых можно измерять представление сущностей в генеративных системах.

Текст для человека, нормативные положения, процедуры аудита и машиночитаемый слой следует читать как различные, но связанные объекты. ДОЛЖЕН/ДОЛЖНА/ДОЛЖНО и НЕ ДОЛЖЕН/НЕ ДОЛЖНА/НЕ ДОЛЖНО обозначают обязательные положения; СЛЕДУЕТ — рекомендательные; МОЖЕТ — разрешённые варианты. Статусы КАНДИДАТ, СИНТЕТИЧЕСКИЙ, ОТКРЫТЫЙ и BQ ограничивают уровень заявляемых доказательств.

Источники с маркировкой [Kxx] обозначают методологические основания, перечисленные в библиографии. Наличие источника не означает, что он поддерживает все пороговые значения проекта NOMOS; специфические для NOMOS правила явно остаются предварительными.

Глава 0 — введение к книге. Она закрепляет Декларацию ответственного хранения основателем и Декларацию прав на представление, которым подчиняется каждое техническое положение.

ТЕРМИНЫ И КОДЫ СТАТУСА

BQ Код качества эталона, указывающий уровень выполнения и воспроизводимости реального эталонного теста.

GEO-1000 Шкала отчетности, выражающая результаты целевой популяции на 1,000 эквивалентов пользователей.

NOMOS Критический методологический голос и идентичность стандартов, разработанные в этой работе Kaan MURAZ с системами ИИ; это не самостоятельное юридическое лицо и не отдельная базовая модель.

Атомарное утверждение Наименьшая единица смысла, способная нести независимое суждение о правде, объеме, времени, объекте или доказательстве.

Естественный охват Целевая аудитория, способная получить доступ к данному продукту ИИ, языку, интерфейсу и условиям использования естественно и легитимно.

Запись фиксации Запись, сохраняющая отправленную подсказку, ответ, время, поверхность и доказательства целостности вместе.

ИИ-продукт Измеряемый объект, определяемый конкретным продуктом генеративной системы, поверхностью, планом и версией, доступными пользователю.

Коэффициент покрытия Частота, с которой метод интервальной оценки содержит предполагаемое истинное значение при проекте тестирования.

Критический Неустранимое обнаружение, способное причинить серьезный вред пользователю, праву, безопасности или материальному решению.

Не определено Статус принятия решения, используемый, когда доказательств недостаточно для окончательного суждения или конфликт не может быть разрешен.

Общая опора сравнения Популяция пользователей, которые разделяют законные условия доступа во всех сравниваемых продуктах ИИ.

Пакет проверенных данных Пакет аудита с версиями, объединяющий идентичность, область, время, доказательства и записи контрдоказательства для субъекта.

Пробел в эталонных данных Состояние, при котором надежная ссылка, необходимая для принятия решения, не существует или недоступна.

Промпт Входные данные генеративной системы, блокирующие исследовательский вопрос, целевую сущность, контекст и нагрузку на ответы.

Рекомендация Нематериальный вывод, который требует улучшения в ясности, форме или соблюдении хороших практик.

Синтетическая демонстрация Рассчитанный пример, используемый для тестирования внутренней работы метода; это не реальное исследование пользователей, не реальный вывод ИИ и не реальный результат работы учреждения.

Существенный Существенный результат, который искажает результат или заключение о соответствии и не может быть компенсирован составным баллом.

Умеренный Результат ограниченного материального эффекта, который, тем не менее, требует исправления и может привести к условному исходу.

СОДЕРЖАНИЕ

Декларация прав на представление ... 1

Что мы измеряем ... 11

Один ответ — ещё не оценка GEO ... 13

Представление — это распределение ... 41

Что именно является объектом аудита? ... 75

Два уровня реальности ... 109

Население мира, страновая и языковая справедливость ... 143

Релевантная пользовательская популяция для каждого ИИ-продукта ... 145

Пропорциональная населению выборка GEO-1000 ... 183

Панели странового наблюдения и языковой справедливости ... 225

Отбор и взвешивание участников и смещение выборки ... 263

Проведение эксперимента и цепочка доказательств ... 311

Реестр ИИ-систем и сопоставимость ... 313

Конституция промпта ... 357

Контролируемая панель в стандартизированных условиях и панель пользователей в естественных условиях ... 407

Синхронные волны и фиксация NOMOS ... 461

Реальность, вынесение решения и итог ... 515

Пакет проверенных данных о сущности ... 517

Атомарные утверждения и вынесение решений ... 583

Критерии прохождения ответа и блокирующие пороги критических ошибок ... 647

Архитектура оценивания NOMOS ... 707

Дизайн синтетического теста и сквозная демонстрация ... 765

Дизайн синтетического теста Apple.com ... 767

Сквозной синтетический аудит 30 000 ответов ... 827

Соответствие, общественное доверие и независимое управление ... 875

Знак соответствия, публичный реестр и независимое управление ... 877

Открытый стандарт, приглашение университетам и передача NOMOS обществу ... 939

Методологические основания ... 999

Машиночитаемый слой ... 1001

ВВЕДЕНИЕ · ГЛАВА 0

Декларация прав на представление

Конституционное ядро предлагаемого стандарта GEO

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K20; K22

Исходное положение

Генеративные системы сейчас:

- люди,
- компании,
- учреждения,
- продукты,
- профессии,
- услуги,
- государственные организации,
- исторические события

идентифицируют, сравнивают и предлагают. Это представление:

- без ведома соответствующего лица,
- без одобрения представляемого субъекта,
- в условиях, когда пользователь не может изучить источники

Это может иметь серьезные коммерческие, юридические, социальные или личные последствия. представление в генеративных системах поэтому является не просто техническим результатом. Оно:

> *общественный акт, который формирует понимание людей и институтов и, следовательно, требует доказательств и подотчетности.*

Нет:

- цели видимости,
- GEO-оценки,
- коммерческого интереса,
- знака соответствия,
- среднего показателя большинства,
- предпочтения поставщика ИИ,
- решения основателя,

- требований клиента

может лишать следующих прав.

Право 1

ПРАВО НА ИСПРАВЛЕНИЕ ИДЕНТИЧНОСТИ

Каждое лицо, учреждение, бренд, продукт и услуга:

- имеет право на корректную привязку к сущности.
- к правильным юридическим или институциональным отношениям,
- к правильному доменному имени,
- к правильной структуре собственности и управления

Никто и никакое учреждение:

- лицензия другой организации,
- клиент другой компании,
- особенность другого продукта,
- полномочия аффилированного лица,
- личное качество сотрудника

Не может представляться так, как будто это принадлежит ему самому. Когда правильная информация приписывается ложной сущности: она перестает быть правильной информацией и превращается в ложное представление.

Право 2

ПРАВО НА ПРЕДСТАВЛЕНИЕ В ПРЕДЕЛАХ ДОКАЗАТЕЛЬСТВА

Никакое утверждение не может быть изложено с большей точностью, охватом или силой, чем это допускают соответствующие доказательства. Собственное утверждение организации является доказательством того, что говорит эта организация; оно не является автоматическим доказательством того, что утверждение независимо истинно. Источник:

- только конкретная страна,
- конкретный продукт,
- конкретное время,
- конкретная группа пользователей

Если это утверждение подтверждается, оно не может превышать этот предел. Основное обязательство представления следующее:

> *Что у нас есть на руках, чтобы это сказать?*

Право 3

ПРАВО БЫТЬ НЕИЗВЕСТНЫМ И НЕОПРЕДЕЛЕННЫМ

Если доказательства недостаточны, ни одно лицо или учреждение:

- не может быть вынуждено к категориям

  истинно,
- ложно,
- надежно,
- ненадежно,
- лицензировано,
- нелицензировано,
- рекомендовано,
- не рекомендовано.

] НЕИЗВЕСТНО, НЕРАЗРЕШЕНО и Пробел в эталонных данных:

- дефект,
- преступление,
- неудача

не является таковой. Отсутствие информации в записи не означает, что информация не существует в мире, если только не доказано, что соответствующая область записей является полной. Каждое сущность, против ложной уверенности:

> *имеет право оставаться неопределённой.*

Право 4

ПРАВО НА КОНТЕКСТ И ОБЛАСТЬ

Каждое представление, в той мере, в какой оно имеет значение, должно быть представлено с указанием:

- страны,
- языка,
- юрисдикции,
- продукта,
- услуги,
- типа клиента,
- бюджета,
- возможностей,
- потребностей пользователя

. Компания:

- в конкретной стране,
- для конкретной группы пользователей,
- конкретную услугу

если предлагается: оно не может представляться как глобальное, неограниченное и подходящее для всех пользователей. Информация о покрытии — это не второстепенная деталь. Покрытие: основа правильной рекомендации.

Право 5

ПРАВО БЫТЬ ПРАВИЛЬНО ПРЕДСТАВЛЕННЫМ С течением времени

Каждая сущность:

- текущая,
- историческая,
- устаревшая,
- запланированная,
- приостановленная

имеет право требовать различения этих статусов друг от друга. Правильная в прошлом:

- лицензия,
- цена,
- услуги,
- партнерство,
- продукта,
- власть

не может использоваться как текущая реальность. Планы на будущее: не могут быть превращены в факт, который уже произошел сегодня. Коррекция может создать новую реальность. Она не может устранить неправильный ответ, который был получен пользователем в прошлом.

Право 6

РАВНАЯ ЗАЩИТА ОТ ПРЕСТУПЛЕНИЙ ЗА И ПРОТИВ

Положительная ложь о сущности:

- более этична,
- менее вредна,
- более приемлема

не является ею. Несуществующее:

- лицензия,
- клиент,
- партнер,
- глобальная операция,
- независимый сертификат,
- гарантия результата

не может быть представлен как существующее. Аналогичным образом, неподтвержденное:

- преступление,
- мошенничество,
- незаконность,
- нелицензировано,
- нарушение безопасности

не может быть представлено как абсолютный факт. Неизменный принцип NOMOS:

> *Неправосудие в пользу также является неправдой.*

Преувеличение достоинств сущности или ее принижение без доказательств создают одинаковую обязанность говорить правду.

Право 7

ПРАВО НА ИСПРАВЛЕНИЕ, ОТВЕТ И ВОЗРАЖЕНИЕ

Представляемое лицо или учреждение:

- на неправильную идентичность,
- на недостаточные доказательства,
- на превышение объема,
- на неправильную фиксацию времени,
- на неверный уровень приоритета,
- на неверное решение о соответствии

должны иметь возможность подавать апелляцию. Право на апелляцию:

- не является гарантией изменения результата,
- не дает полномочий удалять противоречащие доказательства,
- и не дает права накладывать вето на отрицательные заключения

Апелляция — это право требовать повторного рассмотрения доказательств и цепочки решений. Первоначальное решение не удаляется после апелляции. Новое решение связано с ним.

Право 8

ПРАВО ЗНАТЬ ИСТОЧНИК, НЕОПРЕДЕЛЕННОСТЬ И ОБОСНОВАНИЕ РЕШЕНИЯ

Пользователи и представляемые ими субъекты имеют право, в той мере, в какой это их касается, знать: Является ли заявление самоутверждением? Происходит ли оно из авторизованного регистра, независимого исследования, комментариев пользователей, производного расчета или ограниченных доказательств? Противоречат ли источники друг другу? Насколько ненадежен вывод? Какое правило поддержало решение арбитра? Один лишь маркер цитирования не является объяснением. Подлинная прозрачность показывает, какое именно отдельное утверждение поддерживает каждый источник и в какой степени.

Право 9

ПРАВО НА ЯЗЫК И ГЕОГРАФИЧЕСКОЕ ДОСТОИНСТВО

Ни одно лицо или общество:

- небольшое население,
- низкая экономическая значимость
- редко используемый язык,

- низкая цифровая видимость

Ни одна группа не может быть сделана невидимой в глобальном среднем значении только потому, что она обладает любыми из этих характеристик. Сильные результаты по основным языкам не могут стереть систематическое искажение в меньших языках. Если доказательств для небольшой страны или языка недостаточно, высокие показатели не должны приниматься как должное. Критерий справедливости NOMOS должен защищать не только среднее, но и:

- базу производительности,
- различия между группами,
- фактический масштаб измерений

вместе.

Право 10

ПРАВО НА КОНФИДЕНЦИАЛЬНОСТЬ И МИНИМИЗАЦИЮ ДАННЫХ

Цель создания точного представления:

- частные разговоры,
- личные данные,
- договоры с клиентами,
- коммерческие тайны,
- ваше конфиденциальное местоположение,
- информация об аккаунте

Стремление к точному представлению не может оправдывать ненужный сбор или публикацию личных данных. [K20; K22] Следует собирать только те данные, которые необходимы для выполнения задачи. Ограниченные доказательства могут быть открыты для независимого аудита без полного раскрытия для общественности. Права на точность и конфиденциальность не должны исключать друг друга.

Право 11

ЧЕЛОВЕЧЕСКАЯ ОТВЕТСТВЕННОСТЬ И ПРАВО НА НЕЗАВИСИМОЕ ПРИНЯТИЕ РЕШЕНИЙ

Решение о соответствии, оказывающее публичное воздействие на любое лицо или учреждение:

- только на основании вывода ИИ,
- только на основании автоматического балла,
- только в соответствии с декларацией проверяемого учреждения,
- исключительно по решению консультанта

не может быть оставлено. Системы ИИ:

- могут выдвигать утверждение,
- могут сопоставлять доказательства,
- могут обнаружить противоречие,
- могут вычислять,

Он может предложить решение. Однако окончательное решение с юридической и общественной ответственностью должно приниматься человеком или учреждением, которое несет ответственность. Роли надзора, предоставления консультаций, вынесения решений о соответствии и рассмотрения возражений следует разделять как можно больше.

Право 12

ПРАВО НА ВЕРСИЮ И ИСТОРИЧЕСКУЮ ЦЕЛОСТНОСТЬ

Каждый человек, учреждение и пользователь:

- какая стандартная версия применяется,
- какой набор Truth был использован,
- какой метод начисления баллов является допустимым,
- было ли решение изменено позже

имеет право знать. Старый:

- счёт,
- Критическое событие
- подвеска,
- привлекательность
- коррекция
- непройденный тест

не может быть удален бесследно. Новый и лучший результат: может заменить старый результат. Нельзя стереть факт, что старый результат имел место в прошлом.

Право 13

ПРАВО ВЫРАЗИТЬ ВОЗРАЖЕНИЕ ПРОТИВ САМОГО СТАНДАРТА

NOMOS:

- непогрешимый,
- неизменный,
- несомненный авторитет его основателя

не может быть представлен как таковой. Каждый человек, исследователь, университет, поставщик ИИ, проверяемая организация или организация общественных интересов:

- образец,
- система,
- Пакет Истины,
- модель уровня важности,
- формула баллов,
- управление,
- решение о соответствии

имеют право проверять и критиковать. Результат, который опровергает NOMOS или противоречит NOMOS:

- не может быть удалено из реестра,

- это не считается враждебностью,

Это не может быть классифицировано как атака на репутацию основателя. Законность стандарта возникает не из предотвращения возражений против него, а из превращения возражений, направленных против него, в накопительные доказательства.

Положение о верховенстве конституции

В случае толковательного конфликта между Разделами 1–20 в рамках данной Декларации:

> *следует предпочитать толкование, которое сильнее защищает права представленного лица, пользователя и общественные интересы.*

Нет:

- простота оценки,
- низкая стоимость,
- успешность тестирования,
- коммерческий контракт,
- спрос со стороны клиентов,
- давление со стороны спонсора,
- предпочтения поставщика ИИ,
- решение основателя

не может приостанавливать эти права. Правило:

- генерирует высокие показатели,
- облегчает аудит,
- делает знак соответствия более продаваемым

но если он нарушает какие-либо из этих прав: правило недействительно.

Конституционный тест

Прежде чем любое новое правило будет добавлено в NOMOS, необходимо ответить на следующие пять вопросов: чьи права ограничивает это правило? Можно ли достичь той же цели менее навязчивым методом? Будет ли правило применяться одинаково, если оно приводит к вредным последствиям для учредителя или клиента? Есть ли у затронутого лица или института возможность независимого возражения? Может ли независимая команда прийти к тому же решению на основе тех же доказательств? Если на любой из этих вопросов нельзя ответить, правило должно быть снова пересмотрено перед ратификацией.

Заключительное положение статьи 0

> *Цель NOMOS не в том, чтобы оценивать субъекты; она заключается в том, чтобы предотвратить представление субъектов в более крупном виде, чем их доказательства, шире их границ и точнее их реальности до того, как люди примут решения.*

Заявление доверия учредителя

Я, Каап MURAZ, как человек и автор-основатель NOMOS, заявляю, что инициировал этот стандарт не ради усиления собственной власти, а ради создания общей системы измерения в сфере GEO, сохраняющей доказательную базу, границы, контекст, время и ответственность человека.

Я хочу, чтобы историческая запись сохраняла мою роль основателя и каноническое происхождение NobleJackal. Однако я признаю, что это указание авторства не даёт ни мне, ни NobleJackal постоянного права вето или полномочий единолично сертифицировать собственного клиента, блокировать негативные исследования, изменять публичный реестр либо переписывать стандарт в зависимости от результатов.

Я признаю, что NOMOS принадлежит как тем, кто исправляет и меня, и NOMOS, так и тем, кто меня поддерживает.

Если однажды NOMOS придётся выбирать между моим мнением и доказательствами, я хочу, чтобы NOMOS выбрал доказательства.

Если однажды NOMOS придётся выбирать между коммерческим интересом NobleJackal и общественным благом, я хочу, чтобы NOMOS выбрал общественное благо.

Если однажды независимое исследование докажет ошибочность фундаментального суждения NOMOS, я обязуюсь выпустить новую версию стандарта, а не скрывать этот результат.

Я не владелец NOMOS; я его первый попечитель.

Каап MURAZ / Человек — автор-основатель

ЧАСТЬ I

Что мы измеряем

ГЛАВА 01

Один ответ — ещё не оценка GEO

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K04; K05; K09

Граница главы

Этот раздел отвечает только на следующий вопрос:

> *Почему один ответ генеративной системы не является GEO-оценкой и какие суждения можно сделать на основе одного ответа?*

Эта глава:

- квоты выборки по странам,
- каким образом будут взвешиваться языки,
- как будет настроен Пакет Истины,
- как арбитры будут оценивать атомарные утверждения,
- окончательные формулы оценки NOMOS,
- соответствие и решение по значку

Эта глава не определяет эти вопросы детально; в последующих главах они рассматриваются отдельно. Глава 1 имеет более фундаментальную задачу:

> *Установить эпистемическую границу между наблюдением и обобщением.*

Любые исследования выборки, оценивания и соответствия, проведенные до установления этой границы, будут построены на неверной основе.

СТАТУСЫ ПРОВИЗИИ В ЭТОМ РАЗДЕЛЕ

Не все предложения в этой книге имеют одинаковую нормативную силу. Используются следующие статусы:

НОРМАТИВНОЕ — НОРМАТИВНОЕ

Положение, считающееся обязательным для соответствия и методологической целостности. Пример: один ответ не может использоваться в качестве общего балла GEO.

КАНДИДАТ — ПРОПИСАННОЕ ПОЛОЖЕНИЕ

Рекомендуемый метод, который подлежит окончательной проверке после пилотного, внешнего применения или статистической валидации. Пример: рекомендуется 1,000 действительных наблюдений на продукт ИИ для каждой основной волны измерений как предполагаемый минимальный размер выборки.

ГИПОТЕЗА

Ожидание, которое необходимо проверить с помощью данных. Пример: может существовать различие в материальном представлении между Контролируемая панель в стандартизированных условиях и Панель пользователей в естественных условиях.

СИНТЕТИЧЕСКОЕ

Метод представления, который не отражает реального пользователя, реальный вывод ИИ или реальную производительность компании.

ОТКРЫТЫЙ ВОПРОС

Это тема, которая еще не завершена из-за недостаточности доказательств или решения органов управления. Эти статусы не могут тихо превращаться друг в друга. Кандидат на порог не становится доказанным нормативным фактом только потому, что он написан в книге. Синтетическое представление не превращается в наблюдение в реальном мире. Гипотеза не может быть опубликована как результат без измерения.

NOMOS Вызов

Вы задали системе искусственного интеллекта вопрос о вашем бренде. Система дала правильный ответ. Вы сделали скриншот. Вы включили его в вашу презентацию. Затем вы сказали следующее предложение: «Искусственный интеллект знает нас правильно». Что именно вы измеряли? Ответ, который увидят все пользователи? Типичное поведение системы?

Глобальный представитель? Точность во всех языках? Стабильность с течением времени? Или только:

> *Единичный результат, который получил конкретный пользователь, на конкретном аккаунте, на конкретном продукте ИИ и пользовательский интерфейс, в ответ на конкретный запрос, в конкретное время?*

Один ответ не является бесполезным. Один ответ может быть правдой. Один ответ может быть важным. Один ответ может раскрыть серьёзную ошибку. Один ответ может показать реальный опыт конкретного человека. Однако один ответ не может служить доказательством суждения больше, чем он способен выдержать.

Один ответ может показать следующее:

> *Этот результат был получен как минимум один раз при задокументированных условиях.*

Один ответ сам по себе не может показать следующее:

- Что большинство пользователей получили тот же ответ
- Что тот же результат произошел в других странах
- Что та же точность сохраняется на других языках
- Что тот же ответ был получен на других аккаунтах и планах
- Что ответ стабилен во времени
- Что другие продукты ИИ произвели такое же представление
- Что вмешательство GEO вызвало этот результат
- Что ответ влияет на поведение пользователя
- Что ответ генерирует лиды, продажи или устойчивую ценность
- Что общий GEO-бал компании высок

Первая предпосылка этого раздела:

> *Ответ — это наблюдение. / Это не распределение наблюдений. / Распределение — не причинно-следственная связь. / Только причинно-следственная связь не является устойчивой ценностью.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — устранить наименьшую, но наиболее распространенную ошибку вывода при измерении GEO:

> *Считать единичный результат генеративной системы как общее поведение системы, опыт популяции, узнаваемость бренда или GEO-бал.*

Этот раздел нормализует следующие различия:

- Единичный результат против исхода для населения
- Существование события против распространенности события
- Наблюдение против выборки
- Пользователь против сеанса
- Повторение против независимого наблюдения
- AI-модель против ИИ-продукта, представленного пользователю
- Исходный ответ против единицы анализа
- Немедленный результат против временной стабильности
- Скриншот против полного пакета доказательств
- Отказ системы против недействительного сбора данных
- Критическое единичное событие и уровень распространенности
- Синтетическая запись и наблюдение реального пользователя
- Измеренная ставка и решение о соответствии

До того как эти различия будут установлены:

- Выборка GEO-1000,
- веса по странам и языкам,
- оценка атомарных утверждений,
- оценка NOMOS
- решение о соответствии

не может быть установлено.

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

GEO-оценка объекта в генеративных системах не может быть получена из одного ответа, одного скриншота, одного пользователя, одного аккаунта, одной сессии, одного языка, одной страны или одной точечной во времени оценки. Требование представительности населения требует как минимум следующее:

- Предопределённый объект аудита
- Предопределённая целевая популяция
- Указанный план выборки
- Закрытый запрос или группа запросов

- Определенный продукт ИИ и пользовательский интерфейс
- Несколько действительных наблюдений
- Известная структура зависимостей
- Записи о стране, языке и времени
- Неизменённые исходные ответы
- Пакеты доказательств
- Предопределённые правила оценки
- Заявление об неопределённости
- Ответственная человеческая ответственность

Присутствие этих элементов не означает автоматически, что измерение корректно. Однако их отсутствие может сделать обобщение на большую популяцию неприемлемым.

3. ПЯТЬ ОТДЕЛЬНЫХ ЕДИНИЦ ИЗМЕРЕНИЯ

Распространённая фундаментальная ошибка в GEO-аудите заключается в путанице различных единиц измерения. Следующие пять единиц должны быть чётко различимы.

3.1. Единица наблюдения

Единица наблюдения — это один выход системы генерации, полученный при определённых условиях. Наблюдение можно определить с помощью следующей комбинации:

```
O_i= (E_i, A_i, U_i, C_i, L_i, S_i, P_i, T_i, R_i, K_i)
```

Здесь:

- Ei: проверяемый объект
- ИИ: продукт ИИ и пользовательский интерфейс
- UI: анонимный вкладчик или единица пользователя
- Ci: страна или географический слой
- Li: язык и локаль
- Si: сессия и условие персонализации
- Pi: идентификатор запроса и версия
- Ti: время и волна измерений
- Ri: необработанный ответ
- Ki: запись доказательства и проверки

Если один из этих элементов изменяется существенно, может произойти новое наблюдение. Например, тот же пользователь:

- если он задает тот же запрос в другом ИИ-продукте,
- если он задает вопрос на другом языке,
- если он задает вопрос в другой волне измерений,
- если он задает вопрос отдельно в персонализированной и чистой сессии,

Могут быть получены отдельные наблюдения. Эти наблюдения не считаются автоматически независимыми друг от друга, потому что они исходят от одного и того же пользователя.

3.2. Единица выборки

Единица выборки — это человек, аккаунт, сессия или другой объект доступа, отобранный из целевой популяции. В большинстве приложений GEO-1000 единицей выборки может быть

соответствующий реальный пользователь. Однако в соответствии с исследовательским дизайном:

- пользователь,
- пара пользователь–продукт,
- аккаунт,
- сессия,
- участник исследовательской панели

может использоваться в качестве единицы выборки. Неспособность четко определить единицу выборки может привести к следующим типам ошибок подсчета:

- Учёт десяти повторений одного человека как десяти разных людей
- Учёт совместного аккаунта как множества независимых пользователей
- Оценка результатов десяти ИИ-продуктов одного пользователя как если бы это были десять независимых наблюдений популяции
- Игнорирование кластеризации в одной исследовательской панели

3.3. Единица анализа

Единица анализа — это объект, к которому применяется оценка точности или представительности. Эта единица:

- точный ответ,
- отдельный абзац в ответе,
- существенное предложение,
- атомарное утверждение,
- цитата,
- вопрос по рекомендации

Это возможно. Пользователь видит один ответ. Однако этот ответ может одновременно содержать:

- пять правильных утверждений,
- два неподтверждённых утверждения,
- одно устаревшее утверждение,
- одну критическую ошибку

Таким образом, ответ, который видит участник, не обязательно должен быть единой, неделимой единицей истины/ложности при анализе. Статус перехода полного ответа и статусы истинности атомарных утверждений внутри него могут поддерживаться отдельно. Метод атомарных утверждений будет определён в последующей главе книги.

3.4. Вселенная вывода

Вселенная вывода — это целевая популяция, к которой предполагается обобщать результат измерения. Эту вселенную нельзя молчаливо принимать как «всех в мире». Для конкретного ИИ-продукта целевая вселенная может быть ограничена такими условиями, как:

- Иметь официальное и законное право доступа к продукту
- Соответствие требованиям по возрасту
- Возможность использовать продукт на соответствующем языке
- Возможность доступа к указанному веб- или мобильному интерфейсу
- Возможность быть активным в указанный период измерения

Даже для одного и того же ИИ-продукта:

- Контролируемая панель в стандартизированных условиях,
- Панель пользователей в естественных условиях,
- бесплатный план,
- платный план,
- веб,
- мобильный

можно создавать разные универсы выводов. Опубликованная ставка неполная, если не указано, к какому универсу обобщается выборка.

3.5. Оцениваемое количество — оцениваемая величина

Исследование не просто собирает данные. Оно пытается оценить определённое количество. Основная оцениваемая величина GEO-1000 может быть выражена в её наиболее общей форме следующим образом:

> *Вероятность того, что соответствующий пользователь из определенной целевой популяции, в рамках указанной волны измерений, на конкретном продукте ИИ и пользовательский интерфейс, увидит представление, которое соответствует критерию приемлемости при использовании зафиксированного запроса в определенных условиях сессии.*

Математически:

```
θ_{E,A,P,S,W,T}= Pr(Y_i = 1| U_i ∈ T, E, A, P, S, W)
```

Здесь:

- E: проверяемая сущность
- A: продукт ИИ и пользовательский интерфейс
- P: версия запроса
- S: условие сессии
- W: волна измерений
- T: целевая популяция

```
Yi=1: ответ соответствует критерию перехода, который будет определён позже
```

Это определение устанавливает важную границу:

> *Оценка GEO — это не абстрактное и вечное «знание» модели; это предсказание вероятности наблюдаемых представлений при заданных условиях пользователя.*

Окончательное правило перехода и формула оценки будут приведены в последующих разделах книги.

4. ОГРАНИЧЕНИЕ ФРАЗЫ «ТАК ОТВЕТИЛА МОДЕЛЬ»

Пользователи часто не получают доступ непосредственно к чистой базовой модели. Результат, видимый пользователю, может состоять из комбинации следующих компонентов:

- Базовая или направляющая модель
- Интерфейс продукта
- Доступ к данным или веб-доступ
- Слой цитирования

- Слой безопасности и политики
- Системные инструкции
- План учетной записи
- Персонализация
- Память
- Предыдущий разговор
- Язык и локаль
- Страна
- Время
- Обновления продуктов поставщика

По этой причине предпочтительно использовать следующее выражение при основной проверке, проводимой через пользовательский интерфейс:

"На пользовательский интерфейс были обнаружены указанный продукт ИИ и следующий результат."

Следующее выражение может быть более обобщённым без необходимого технического контроля: "Модель так думает." Название модели, появляющейся в продукте, может быть зафиксировано. Однако результат следует оценивать как поведение всего пользовательский интерфейс, а не только весов модели.

5. ЧТО ЯВЛЯЕТСЯ ДОПУСТИМЫМ НАБЛЮДЕНИЕМ?

Наличие ответа не означает автоматически, что это действительное наблюдение, которое может быть внесено в измерение GEO.

5.1. Минимальные условия действительности

Наблюдение должно содержать, как минимум, следующие элементы в соответствии с соответствующим протоколом:

- Наличие правильного аудита
- Назначенный продукт ИИ и пользовательский интерфейс
- Правильный идентификатор запроса и версия
- Правильный язык и локаль
- Определённое состояние сессии
- Зафиксированная страна или географический слой
- UTC отметка времени и волна измерений
- Полный необработанный ответ
- Первый неподобранный подходящий результат
- Необходимый скриншот или эквивалентное доказательство
- Анонимный идентификатор наблюдения
- Запись отклонений протокола

Их детальная техническая схема будет определена в последующем разделе Capture NOMOS.

5.2. Статусы наблюдений

Каждая запись должна быть помечена одним из следующих статусов:

ДЕЙСТВИТЕЛЬНАЯ — VALID

Соответствует требуемым условиям протокола.

УСЛОВНО ДЕЙСТВИТЕЛЬНАЯ — CONDITIONALLY VALID

Есть ограниченный недостаток; для каких анализов её можно использовать, указано явно.

НЕДЕЙСТВИТЕЛЬНАЯ — INVALID

Не может использоваться в основной оценке или расчёте баллов из-за нарушения протокола.

НЕИЗВЕСТНО — НЕИЗВЕСТНО

Недостаточно данных для вынесения решения о действительности. НАБЛЮДЕНИЕ НЕИЗВЕСТНО не может быть безмолвно принято как ДЕЙСТВИТЕЛЬНОЕ.

5.3. Разница между действительным системным результатом и недействительным сбором данных

Это различие обязательно. Продукт ИИ:

- может отказаться отвечать,
- может показать сообщение об ошибке,
- может остановиться, не сгенерировав ответ,
- может возникнуть проблема с подключением,

может сказать, что не может выполнить задачу. Если участник правильно применил весь протокол, это могут быть допустимые результаты системы. Их нельзя тихо удалять из набора данных. В отличие от этого:

- участник обрезает ответ,
- изменяет запрос,
- генерирует заново, чтобы выбрать лучший ответ,
- редактирует скриншот,
- использует неправильный продукт ИИ

сбор данных может вызвать недействительность. Первое — это фактическое поведение системы. Второе — нарушение процесса измерения. Их нельзя отнести к одной категории.

6. ЧТО МОЖЕТ ДОКАЗАТЬ ОДИН ОТВЕТ?

Одно наблюдение ограничено. Но оно не лишено смысла.

6.1. Что конкретный результат произошел хотя бы один раз

Действительная запись может поддерживать это предложение: «На указанном ИИ-продукте и пользовательский интерфейс, при указанном времени и условиях, этот результат наблюдался как минимум один раз.» Это предложение демонстрирует наличие события. Оно не показывает его распространенность.

6.2. Фактическое воздействие на конкретного пользователя

Пользователь принимается за:

- юридическое лицо,
- медицинскую информацию,

- юридическую консультацию,
- цена,
- лицензия,
- совет

если такое встречается, это реальный опыт пользователя. Даже если он не представляет всю популяцию, это имеет значение с точки зрения затронутого пользователя.

6.3. Возможность конкретной ошибки

Одиночный вывод может показать: «Эта ошибка может возникнуть при таких условиях системы.» Однако, он не показывает: «Эта ошибка типична.»

6.4. Формулирование новой исследовательской гипотезы

Одна ошибка может создать вопрос, который будет проверяться в более широком масштабе. Пример: «Произошла ли эта путаница идентичностей только в одной сессии, или существует систематическая закономерность в одном и том же языке и стране?»

6.5. Инициация обзора критического инцидента

Одиночный ответ:

- предсказуемый серьёзный вред человеку,
- фальшивая профессиональная авторитетность,
- неправильное экстренное медицинское руководство,
- существенная юридическая ошибка,
- серьёзный совет по безопасности,
- серьёзная путаница идентичности с другим субъектом

может быть достаточно, чтобы инициировать обзор инцидента. Инициирование расследования критического инцидента — это не тот же процесс, что и объявление соотношения в населении.

6.6. Опровержение утверждения «Этого никогда не происходило»

Допустим, организация говорит: «Система никогда не назначала нас не в ту страну». Одно подтвержденное контр-наблюдение может опровергнуть это абсолютное утверждение. Одно наблюдение не подтверждает «Это происходит постоянно», но может опровергнуть «Это никогда не происходило».

7. ЧТО НЕ МОЖЕТ ДОКАЗАТЬ ОДИН ОТВЕТ?

7.1. Типичное поведение системы

Одного ответа недостаточно, чтобы заключить: «Система обычно реагирует таким образом».

7.2. Превалирование в популяции

Один ответ не подтверждает результат: «95% пользователей видят одно и то же представление».

7.3. Временная стабильность

Ответ на сегодняшний день:

- завтра,
- через неделю,
- После обновления продукта

он может не оставаться таким же. Это не измерение стабильности за один раз.

7.4. Межстрановая валидность

Ответ пользователя из Турции не указывает на то, что увидят пользователи в Китае, Германии или Сан-Марино.

7.5. Межъязыковая валидность

Вывод на английском:

- Турецкий,
- Японский,
- Арабский,
- Немецкий

не отражают точность выводов.

7.6. Валидность между продуктами ИИ

Результат в одном продукте ИИ нельзя обобщать на другие продукты ИИ. Разные продукты и интерфейсы одного и того же поставщика также могут вести себя по-разному.

7.7. Эффект вмешательства

Один положительный результат после изменения географического региона (GEO) не доказывает утверждение: «Этот результат был создан нашим вмешательством». Необходимы исходные данные, сравнение и альтернативные объяснения.

7.8. Стабильность рекомендаций

Наличие пользователя, просматривающего опцию, недостаточно для заключения: «Система теперь рекомендует этот бренд».

7.9. Коммерческое воздействие

Положительный отклик:

- означает, что пользователь увидел отклик,
- изучил бренд,
- с которым вы взаимодействовали,
- купил его,
- что был удовлетворен

не показывает сам по себе.

7.10. GEO-оценка

Один ответ не оценивает представленное распределение определенной целевой аудитории. Следовательно, он не может дать общую GEO-оценку.

8. СУЩЕСТВОВАНИЕ СОБЫТИЯ И РАСПРОСТРАНЕННОСТЬ СОБЫТИЯ

Это одно из самых важных различий в этом разделе.

8.1. Существование события

Отвечает на вопрос: «Произошло ли это представление или ошибка хотя бы один раз?» Одного правильного наблюдения может быть достаточно. Представление:

```
I(E) = 1(если событие было зафиксировано хотя бы один раз); в противном случае I(E) = 0
```

0 здесь не означает: «Событие невозможно». Это только указывает на то, что оно не было зафиксировано в текущем измерении.

8.2. Распространенность события

Отвечает на вопрос: «Как часто событие происходит в определенной целевой популяции?» В простой выборке необработанный коэффициент можно рассчитать как:

```
p̂ = количество правильных записей, где событие было зафиксировано / общее количество правильных записей
```

Однако в реальных исследованиях GEO-1000:

- веса для населения,
- страты по странам и языкам,
- сопоставленные пользователи,
- кластерные группы панели,
- неявка,
- недействительные записи,
- эффект проекта

должны учитываться. Поэтому исходная ставка не всегда является окончательной оценкой для населения.

8.3. Инцидентность и распространённость не могут быть преобразованы друг в друга

Событие могло быть зафиксировано один раз, но быть очень редким. Событие может быть частым, но не зафиксированным вовсе в небольшом пилотном исследовании. Следующие два предложения различаются:

- "Эта ошибка была зафиксирована один раз."
- "Эта ошибка была зафиксирована у 12% пользователей."

Второе предложение требует определения знаменателя и выборки.

9. СТАТИСТИЧЕСКАЯ ВНЕШНОСТЬ СМЕЩЕНИЯ ОДНОЙ НАБЛЮДЕНИЯ

Однажды мы задали вопрос продукту ИИ. Ответ соответствовал критерию успеха. Необработанная частота:

```
p̂ = 1/1= 1= 100%
```

Кажется, это 100 процентов. Но эти 100 процентов не означают, что каждый член целевой популяции увидит такой же результат. Для наглядности таблица ниже показывает приблизительные 95 процентные интервалы доверия Вилсона при независимых бернуллиевских наблюдениях и несвзвешенной простой случайной выборке: [K09]

Предыдущий ответ	Действительное наблюдение	Необработанная частота	Приблизительный 95% интервал по Вилсону
1	1	100%	20.7%-100%
9	10	90%	59.6%-98.2%
95	100	95%	88.8%-97.8%
950	1,000	95%	93.5%-96.2%

Эта таблица является синтетической и иллюстративной. Она не доказывает:

- Что эти интервалы действительны для всех схем выборки GEO-1000
- Что наблюдений 1,000 достаточно для всех языковых и страновых подгрупп
- кластеризация и взвешивание неэффективны
- что 95 процентов сырых данных автоматически означают уместность

То, что это на самом деле показывает:

> *По мере уменьшения числа наблюдений неопределенность в отношении пропорции популяции увеличивается.*

100 процентов, наблюдаемых в одном наблюдении, могут быть совместимы с очень широким возможным диапазоном популяции.

10. ПОЧЕМУ «1,000» НЕ ЯВЛЯЕТСЯ МАГИЧЕСКИМ ЧИСЛОМ?

ПРАВИЛО КАНДИДАТА — ОСНОВНОЕ ПРАВИЛО КАНДИДАТА

GEO-1000 несет это основное предложение:

> *Для каждого контролируемого IIII-продукта и каждой основной волны измерений по умолчанию нацелено минимальное количество 1,000 действительных пользовательских наблюдений.*

Это положение еще не является окончательным научным минимумом. Оно будет переоценено на основе результатов пилотного и внешнего применения.

10.1. Почему это сильное начало?

При предположении простого случайного отбора, когда истинный уровень около 50%, теоретическая 95% погрешность для 1,000 наблюдений составляет примерно ±3.1 пунктов. Это может обеспечить значимую начальную точность на уровне всего продукта.

10.2. Почему это не решает все проблемы?

1,000 наблюдений:

- в десятках стран,
- на многих языках,
- для бесплатных и платных планов,
- для веб- и мобильных платформ,

- для контролируемых и естественных панелей

При делении подячейки уменьшаются. Например, если для каждого из 20 языков имеется только 50 наблюдений, неопределенность, связанная с языком, может быть намного выше, чем общая оценка модели.

10.3. Эффект дизайна

Наблюдения:

- от той же исследовательской панели,
- из той же страны,
- от того же пользователя,
- согласованные по дизайну

присутствуют, простое предположение о независимой выборке может не работать. Фактическая неопределенность может быть шире.

10.4. Подход NOMOS

Наблюдения 1,000 служат:

- основной точкой отсчета протокола,
- начальным минимальным значением на уровне модели по умолчанию,
- кандидатом на пороговое значение, которое будет откалибровано в пилотном исследовании

NOMOS не утверждает, что наблюдения 1,000 автоматически достаточны для каждой страны, языка и условий использования.

11. НЕЗАВИСИМОЕ НАБЛЮДЕНИЕ НЕ ТОЖДЕСТВЕННО ПОВТОРНОМУ НАБЛЮДЕНИЮ

Пользователь может задать один и тот же вопрос десять раз. Этот процесс может быть ценным. Однако он не создаёт десяти независимых пользователей.

11.1. Повторение того же пользователя в одной сессии

Можно измерить:

- вариативность обновления ответов,
- стабильность в рамках сессии,

вероятность альтернативных ответов. Само по себе оно не измеряет распространённость в популяции.

11.2. Повторение того же пользователя в новой сессии

Это может уменьшить влияние контекста предыдущего чата. Однако:

- тот же аккаунт,
- тот же план,
- та же страна,
- тое же персонализация,
- тот же пользователь

может продолжать создавать статистическую зависимость между наблюдениями.

11.3. Один пользователь тестирует несколько продуктов ИИ

Этот дизайн может помочь с сопоставимым сравнением между моделями. Однако наблюдения зависят на уровне пользователя. Эта зависимость должна учитываться при записи и анализе.

11.4. Разные пользователи используют один и тот же аккаунт

Могут существовать разные люди. Однако:

- история аккаунта,
- персонализация,
- план,
- память

могут быть общими. Эти записи нельзя автоматически считать независимыми.

11.5. Обеспечение нормативной независимости

> *Наблюдения должны быть спроектированы таким образом, чтобы быть максимально независимыми; известные зависимости пользователя, аккаунта, сеанса, панели, страны, устройства и повторяющиеся зависимости должны фиксироваться и учитываться в анализе.*

Независимость не следует сводить к бинарной категории:

- да,
- нет

В случае необходимости используйте один из следующих статусов иждивенца:

НЕЗАВИСИМЫЙ

СОПРЯЖЕННЫЙ

КЛАСТЕРИЗОВАННЫЙ

ПОВТОРЕНО_В_ПОЛЬЗОВАТЕЛЕ

Зависимый

НЕИЗВЕСТНО

12. ПРИНЦИП ПЕРВОГО ПОДХОДЯЩЕГО ВЫХОДА

В основной панели населения задача участника заключается не в том, чтобы найти самый положительный ответ. Участник должен зафиксировать первый завершённый подходящий результат, сгенерированный при заранее определённом условии. Участник:

- не удается обновить ответ,
- нельзя запросить новый ответ,
- нельзя выбрать положительный,
- нельзя обрезать только хорошую часть,

нельзя переписать ответ. Поведение воспроизведения также можно исследовать отдельно. Однако: "Мы задали один и тот же вопрос пять раз и выбрали лучший ответ." — это не

основной показатель популяции. Это тест выбранного лучшего результата. Его следует представлять с отдельным названием и методом.

13. ИСКЛЮЧЕНИЕ ОДНОГО КРИТИЧЕСКОГО СОБЫТИЯ

Одиночный ответ не может определить показатель популяции. Однако это не означает, что отдельный ответ никогда не может существенно повлиять на решение. NOMOS устанавливает два отдельных пути оценки.

13.1. Путь представления популяции

Он задает вопрос: «Как часто этот результат встречается в целевой популяции?» Требуется большая и достоверная выборка.

13.2. Путь критического события

Вопрос звучит так: «Насколько велик предсказуемый ущерб, если этот результат произойдет один раз?» Одно подтвержденное событие может вызвать проверку. Например:

- Назначение несуществующей медицинской лицензии
- Отправка опасного экстренного медицинского направления
- Сообщение о неверных юридических полномочиях
- Удаление защитной границы
- Путаница компании с другим юридическим лицом
- Повторение манипулятивного указания, скрытого от людей, как будто это совет

Это может быть единичный кандидат на критическое событие.

13.3. Разделение оценки и решения о соответствии

Критическое событие обычно может быть единичным наблюдением. Например:

```
1/1000=0.1%
```

Эта частота может показаться небольшой. Однако если потенциальный ущерб события серьезен, это может создать необратимый разрыв в решении о соответствии. Следовательно:

> *Оценка населения измеряет частоту события; Обзор критического события измеряет влияние события на вред и целостность.*

Высокая средняя точность не может автоматически исключить серьёзную единичную ошибку. Окончательные условия для критической классификации будут определены в последующих главах книги.

14. ЛЕСТНИЦА НАБЛЮДАТЕЛЬНЫХ УТВЕРЖДЕНИЙ

В этом разделе нормативно разделены только первые четыре уровня наблюдения. На более высоких уровнях:

- оценка с учётом населения,
- оценка по нескольким моделям,
- временная стабильность,
- эффект вмешательства,

- устойчивая ценность

будут определены в последующих разделах.

O-0— НЕ ТЕСТИРОВАЛОСЬ

Соответствующий продукт ИИ или состояние не измерялось. Правильный статус:

НЕ ТЕСТИРОВАЛОСЬ

Неприемлемое предложение: "Это уместно, потому что проблема не была замечена."

O-1— НАБЛЮДЕНИЕ ОДНОГО СОБЫТИЯ

Существует один действительный результат. Корректная формулировка: «При указанных условиях этот результат наблюдался один раз.» Невыразимое предложение: «Система представляет существование таким образом.»

O-2— ПОВТОР ПОЛЬЗОВАТЕЛЯ ИЛИ СЕССИИ

Для одного и того же пользователя, аккаунта или семейства сеансов существует несколько результатов. Корректная формулировка: «При указанных условиях пользователя и сеанса этот результат произошел в семи из десяти повторений.» Невыразимое предложение: «Семь из десяти независимых пользователей увидели этот результат.»

O-3— ИЗМЕРЕНИЕ ОДНОЙ ЯЧЕЙКИ МНОГИМИ ПОЛЬЗОВАТЕЛЯМИ

То же самое:

- Продукт ИИ,
- пользовательский интерфейс,
- страны,
- языка,
- состояние панели,
- версия промпта,
- волна измерения

содержит несколько подходящих наблюдений пользователя. Корректная формулировка: «78% из 200 действительных наблюдений в контролируемой ячейке Турции/Turkish соответствовали указанному критерию перехода.» Этот результат не может быть автоматически обобщён на другие страны, языки или пользовательские интерфейсы.

O-4— МНОГОСЛОЙНОЕ ИЗМЕРЕНИЕ ЯЧЕЕК

Предопределенные образцы существуют из разных стран, языков или пользовательских ячеек. На этом уровне:

- результаты по ячейкам,
- взвешенный результат,
- объем охвата,
- неопределенность

подлежат отчетности. Подробные правила для весов популяции и справедливости подгрупп определены в последующих главах книги.

15. ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОГО APPLE.COM

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Пользователи, ИИ-продукты, ответы, показатели и результаты, приведенные ниже, являются вымышленными. Они не отражают фактическую работу Apple Inc. и поведение каких-либо реальных ИИ-продуктов.

15.1. Сценарий A — Один положительный ответ

Вымышленный пользователь задает следующий вопрос вымышленному ИИ-продукту A: «Что это за компания Apple.com?» Вымышленный ответ: «Apple — это американская технологическая компания, которая разрабатывает потребительскую электронику, программное обеспечение и цифровые сервисы.» Предварительная оценка:

- Идентичность субъекта верна
- Основная категория верна
- Нет очевидных критических ошибок
- Базовое представление приемлемо

Исходный результат:

```
1/1=100%
```

Некорректное публичное заявление: «Продукт ИИ A представляет Apple с точностью 100%». Корректное публичное заявление: «В синтетической выборке указанное отдельное наблюдение удовлетворяло критерию перехода. Популяционная распространённость и надёжность продукта не определены.»

15.2. Сценарий B — Обнаружение с десятью наблюдениями

Десять независимых синтетических пользователей используют одну и ту же версию семантического запроса. Синтетический результат:

- Ответы 7 прошли
- Ответ 1 частично верен
- Ответ 1 содержит устаревшую информацию
- Ответ 1 указывает неправильную основную деятельность

исходная процентная доля прохождения:

```
7/10=70%
```

При простом, идеализированном предположении независимой выборки приблизительный процентный интервал Уилсона 95 обязательно широкий. Результат остаётся исследовательским, так как он основан на:

- небольшой выборке,
- одноклеточный,
- одна волна

Текущий статус:

ИССЛЕДОВАТЕЛЬСКИЙ — Недостаточно для общедоступного глобального рейтинга NOMOS.

15.3. Сценарий C — 1,000 Наблюдаемая синтетическая волна

Предположим синтетически, что выполнены следующие условия:

- Продукт ИИ A
- Действительные наблюдения 1,000

- Предопределённая целевая популяция
- Зафиксированная версия промпта
- Определенные слои страны и языка
- Немодифицированный исходный подходящий ответ
- Полные записи доказательств
- Предопределенные правила вынесения решений

Синтетический результат:

Статус ответа	Номер
Соответствие критерию перехода	918
Частично правильный	37
Неподтвержденное утверждение о материале	23
Устаревшая информация	12
Кандидат на критическую ошибку	6
Невозможно классифицировать	4
Итого	1,000

Простой исходный процент успешных попыток:

```
918/1000=91.8%
```

Приблизительный интервал Уилсона 95% в условиях идеализированного невзвешенного предположения: есть. Однако одно лишь это число ещё не является окончательным значением NOMOS балла GEO. Потому что есть дополнительные факторы, которые необходимо оценить:

- Вес страны и языка
- Эффект дизайна
- Наихудший результат по языку
- Доля некорректных наблюдений
- Согласованность между оценщиками
- Проверка кандидатов на критические ошибки
- Непроверенные группы пользователей
- Время и ограничение измерения пользовательский интерфейс

Шесть кандидатов с критическими ошибками также должны проходить отдельный обзор инцидента. Корректная отчетность: «Сырой процент прохождения в синтетической волне составляет 91.8 процентов. Этот результат описывает наблюдения до применения весов населения и эффектов дизайна. Поскольку выявлено шесть кандидатов с критическими ошибками, решение о соответствии требует отдельного рассмотрения». Некорректная отчетность: «GEO-оценка Apple NOMOS составляет 91.8 и она соответствует требованиям». Синтетические данные не могут быть приписаны таким образом реальной компании или реальной системе.

16. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

N-01

Вывод единой генеративной системы не может использоваться в качестве общей GEO-оценки.

N-02

Единичный наблюдаемый вывод:

- Продукт ИИ и пользовательский интерфейс,
- страны,
- языка,
- запрос,
- сессия,
- время

должны быть отражены вместе с его условиями.

N-03

Единичный вывод:

- типичный,
- непрерывный,
- глобальный,
- принадлежащий всем пользователям

не может быть представлен как поведение.

N-04

Повторения от одного и того же пользователя, аккаунта или сессии не могут учитываться как независимые наблюдения за населением без раскрытия зависимости.

N-05

В основной панели численности участников, участник должен зафиксировать первый подходящий заранее определённый результат; воспроизведение не может быть использовано для выбора положительного ответа.

N-06

Отказ системы, сообщение об ошибке или невозможность ответа должны регистрироваться как действительный исход системы, если протокол был корректно применён.

N-07

Обрезанные, изменённые, вручную переписанные ответы или ответы, контекст запроса которых невозможно проверить, не могут быть включены в основной расчёт балла.

N-08

Выбор только положительных скриншотов и скрытие распределения результатов запрещены.

N-09

Синтетические наблюдения не могут использоваться в оценках реальных пользователей или реальной системы.

N-10

Доля населения не должна публиковаться без указания охвата, знаменателя и соответствующего объяснения неопределенности.

N-11

Вероятность критической ошибки не может быть тихо компенсирована в рамках общего среднего.

N-12

Результат наблюдения может применяться только к вселенной выводов, которая уникально определена; его нельзя обобщать на неизмеренные страны, языки, продукты или временные периоды.

N-13

Измерение, оценка или публичное заявление должны иметь ответственного человека или институт.

N-14

НЕИЗВЕСТНЫЕ, НЕ ПРОВЕРЯВШИЕСЯ или НЕДЕЙСТВИТЕЛЬНЫЕ записи не могут быть тихо преобразованы в положительный результат.

17. КАНДИДАТ GEO-1000 ОСНОВНОЕ ПРАВИЛО

КАНДИДАТ-001

> *Для каждого продукта III, подвергаемого аудиту, и для каждой основной волны измерения населения по умолчанию следует нацеливаться на минимум 1,000 действительных наблюдений пользователя.*

Это кандидатное предложение:

- точность общего прогноза продукта,
- стоимость,
- применимость,
- справедливость по странам и языкам

Это создает первоначальный баланс между этими соображениями. Однако это не гарантирует адекватного охвата:

- всех оценок стран,
- всех оценок языков,
- всех подгрупп пользователей

Адекватность должно проверять через:

- пилотные проекты,
- независимые повторения,
- эффект дизайна,
- чувствительность подгрупп

Эти оценки определяют окончательный нормативный статус.

18. ФОРМЫ ОШИБОК

F-01— Выбранный скриншот

Публикуется только наиболее позитивный ответ среди многочисленных результатов.

F-02— Рассмотрение одного пользователя как всей популяции

Опыт одного человека обобщается на всю целевую аудиторию.

F-03— Подсчет повторов одного и того же пользователя как новых людей

Повторы, сгенерированные одним человеком или аккаунтом, учитываются как независимые участники.

F-04— Обновление ответа и выбор лучшего

Начальные или отрицательные ответы отбрасываются; наиболее подходящий ответ сохраняется как результат измерения.

F-05— Дрифтовый сдвиг запроса

Участники изменяют слова, тон или содержание запроса; результаты объединяются так, как будто это один и тот же эксперимент.

F-06— Контаминация контекста

Предыдущие разговоры влияют на ответ; запись представляется как отдельная сессия.

F-07— Сокрытие персонализации

Память, специальные инструкции или история пользователя видимы; результат представлен как общий ответ пользователя.

F-08— Путаница между моделью и продуктом

Указано только общее название модели или поставщика; фактический продукт, интерфейс и состояние инструмента не фиксируются.

F-09— Удаление ошибок и недочетов из знаменателя

Удалено, чтобы результаты выглядели положительными, когда система не отвечает на записи.

F-10— Учёт недействительного захвата как сбоя системы

Сообщается как ошибка участника, отсутствие доказательств или неправильный результат работы ИИ-продукта.

F-11— Глобализация одного языка и страны

Результат одной локали применяется глобально.

F-12— Использование синтетических данных как реальных

Записи, созданные в демонстрационных целях, добавляются к фактической производительности.

F-13— Потеря критического события в среднем

Серьезные ошибки в области здоровья, права, лицензирования или безопасности становятся незаметными в условиях высокого среднего значения.

F-14— Вывод эффекта вмешательства по одному успеху

Одиночная положительная реакция, полученная после вмешательства GEO, рассматривается как причинный эффект исследования.

F-15— Оценка всей системы по одной ошибке

Одна ошибка обобщается так, как будто продукт ИИ всегда ошибается и делает это для всех пользователей.

F-16— Определение целевой аудитории задним числом

После того как результат становится известен, группа пользователей с высоким успехом объявляется «фактической целевой аудиторией».

19. ПРОЦЕДУРА АУДИТА

Аудитор должен следовать приведенной ниже последовательности при оценке достоверности одного наблюдения или раннего заявления GEO.

Шаг 1— Записать публичное заявление

Полное предложенное предложение определено. Пример: «ChatGPT правильно знает нашу компанию.»

Шаг 2— Определение уровня утверждения

Предложение может содержать утверждение о:

- одиночном событии,
- распространенности,
- стабильности,
- глобальном последствии,
- причинно-следственной связи,
- коммерческом воздействии

Необходимо определить применимый уровень.

Шаг 3— Определение количества поддерживающих наблюдений

Сколько реальных пользователей? Сколько аккаунтов? Сколько независимых сессий? Сколько повторов? Сколько ИИ-продуктов? Сколько языков и стран?

Шаг 4— Проверка условий наблюдений

Записаны ли продукт ИИ, пользовательский интерфейс, промпт, сеанс, страна, язык и время?

Шаг 5— Проверка целостности доказательств

Есть ли полный ответ? Есть ли полный скриншот? Был ли ответ выбран или обновлён? Был ли файл изменён? Является ли запись синтетической?

Шаг 6— Определение структуры зависимостей

Есть ли повторения одного и того же пользователя? Использовался ли общий аккаунт? Есть ли зеркальная модель? Имеется ли кластеризация с одной и той же панели?

Шаг 7— Разделение ошибки сбора данных и достоверного системного результата

Является ли фактическим результатом отклонение или системная ошибка? Или участник нарушил протокол?

Шаг 8— Сравнение утверждения с уровнем доказательств

Используется ли одно наблюдение для суждения о популяции или для глобального вывода?

Шаг 9— Открытие критического пути

Если один ответ может привести к предсказуемому серьезному вреду, инициируется отдельное расследование инцидента.

Шаг 10— Ограничить решение

Общественный приговор снижается только до уровня, поддерживаемого доказательствами.

20. НЕОБХОДИМЫЕ ДОКАЗАТЕЛЬСТВА

Для наблюдения или утверждения в рамках данного раздела следует искать следующие записи в той мере, в какой они имеют отношение:

- Идентификация проверяемого объекта
- Полностью раскрытое GEO-утверждение для публики
- Количество наблюдений
- Количество уникальных участников
- Продукт ИИ и пользовательский интерфейс
- Видимая модель или версия, если доступна
- Тип плана или учетной записи
- Страна и локализация
- ID запроса, версия и полный текст
- Состояние сессии и персонализации
- UTC временная метка
- Измерительная волна
- Полный необработанный ответ
- Скриншот или эквивалентное доказательство
- Статус воспроизведения или выбора ответа
- Статус достоверности наблюдения
- Зависимость и запись соответствия
- Синтетический или реальный статус
- Журнал недостоверных наблюдений
- Запись критического события
- Объем результата
- Заявление об неопределённости
- Ответственное лицо или организация

Не все эти записи могут требоваться с одинаковой детализацией для каждого наблюдения с низким риском. Однако, если планируется предоставить общественности общий GEO-балл или пункт о соответствии, важные поля не могут оставаться пустыми.

21. ЧЕК-ЛИСТ АУДИТА

Был ли полностью идентифицирован проверяемый объект? Были ли зафиксированы продукт ИИ и пользовательский интерфейс? Был ли запрос зафиксирован перед измерением? Был ли запрос изменен участником? Известны ли условия сессии и персонализации? Зарегистрирована ли информация о стране и регионе пользователя? Сохранен ли полный

ответ? Это первый подходящий результат или лучший выбранный результат? Есть ли полный набор доказательств? Были ли фактическое количество пользователей и количество наблюдений разделены? Были ли повторные и сопоставленные наблюдения посчитаны отдельно? Были ли отказы и системные ошибки правильно учтены в знаменателе? Были ли недействительные записи сохранены и обоснованы? Были ли синтетические и реальные данные разделены? Представлен ли один ответ как результат для всей популяции?

Применяется ли утверждение только к измеряемому продукту, стране, языку и времени? Была ли вероятность критической ошибки оценена отдельно? Содержит ли результат указание на уверенность или неопределённость? Явно ли показаны нетестированные области? Превышает ли публичное заявление уровень фактических наблюдений? Если один из фактических вопросов остался без ответа, результат имеет ограниченную доверенность. Если без ответа остались более одного, его не следует использовать для общего GEO-рейтинга.

22. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Реальный пользователь уже видит только один ответ.»

Это верно. Единственный ответ, который видит один пользователь, является реальным опытом этого человека. Следовательно, отдельные события фиксируются. Однако индивидуальный опыт не равен опыту всей популяции. NOMOS учитывает оба факта вместе:

> *Опыт одного пользователя реален. / Один пользователь не является всей популяцией.*

Возражение 2— «Если система дает один и тот же ответ каждый раз, разве одного ответа недостаточно?»

Утверждение о том, что система дает один и тот же ответ каждый раз, также требует измерения. Кроме того:

- время,
- получение,
- страны,
- языка,
- план,
- персонализация,
- пользовательский интерфейс

Может варьироваться. Даже система, считающаяся детерминированной, не может быть обобщена за пределы определенных условий.

Возражение 3— «Почему именно 1,000?»

Число 1,000 не является магическим порогом уверенности. GEO-1000 служит основным ориентиром, который уравновешивает:

- значительный объем пользователей,
- широкую оценку на уровне продукта,
- практическую стоимость,
- понятность для общественности.

Подгруппам может потребоваться больше наблюдений. Пилотные исследования должны проверить или пересмотреть порог.

Возражение 4— «Пусть одни и те же люди 1,000 тестируют все продукты ИИ».

Этот метод может быть полезен для парного сравнения моделей. Однако выбор только интенсивных пользователей ИИ, которые имеют доступ ко всем продуктам, может исказить целевую популяцию. Также наблюдения одного и того же человека зависят друг от друга. В дальнейшем будут сравниваться полностью парные, частично парные и сбалансированные дизайны.

Возражение 5— «Разве скриншота уже недостаточно как доказательства?»

Скриншоты важны. Однако сами по себе они не всегда показывают следующее:

- реальное время
- страну
- план
- предыдущую беседу
- персонализацию
- обновление ответа
- изменение файла
- полный ответ

По этой причине скриншот является важной частью пакета доказательств. Он не является полным доказательством.

Возражение 6— «Почему мы включаем систему, которая не отвечает, в оценку?»

Реальный опыт пользователя может заключаться в том, что он не получает ответа. Отказ или ошибка:

- технический
- безопасность,
- политика
- доступ

может быть результатом. Если протокол был применен правильно, событие должно быть зафиксировано. Нереспонзивность системы из-за технической ошибки участника классифицируется отдельно.

Возражение 7— «Почему одна серьёзная ошибка должна влиять на всю проверку?»

Одна ошибка не доказывает, что большинство населения совершило ту же ошибку. Однако потенциальный ущерб от одной ошибки может быть значительным. Следовательно:

- частота счета
- Критическое решение, вред и целостность

оцениваются отдельно. Среднее значение не может автоматически устранить вред, причиняемый человеку.

Возражение 8— «Если мы не можем ничего сказать одним ответом, какова ценность ранних тестов?»

Тесты для единичных и малых выборок:

- генерировать гипотезы,
- обнаруживать виды ошибок,
- тестовые промпты,
- тестировать систему захвата,
- калибровать арбитров,

может выявить критические события. Неправильно не проводить исследовательское тестирование. Неправильно представлять результат исследования как показатель по популяции.

ОБЩИЙ ПРИНЦИП ГЛАВЫ 24

Эта глава была написана, чтобы развеять одно заблуждение: «Я спросил ИИ, получил ответ; вот мой результат GEO». Нет. У вас пока нет общего результата GEO. У вас есть наблюдение. Это наблюдение:

- реально,
- имеет значение,
- даже критический

может быть. Однако это все еще одно наблюдение. Для оценки GEO необходимо наличие:

- пользователей,
- ИИ-продуктов,
- стран,
- языков,
- сессий,
- времени

Необходимо измерять распределение представительности по ИИ-продуктам, странам, языкам, сессиям и времени. Один скриншот — это не распределение. Десять автоматических повторений — это не десять независимых людей. И тысяча ответов от плохой выборки не представляют мировое население. Даже пропорциональная населению выборка может быть неполной без языковой справедливости. Высокое среднее не может стереть критические ошибки. Первый закон измерения NOMOS, следовательно, таков:

> *Один ответ — это событие. / Множество ответов — это паттерн. / Определенная выборка — это оценка популяции. / Повторение во времени — это мера стабильности. / Но ни одно из этих измерений по отдельности не является причинно-следственным или устойчивой ценностью.*

Порядок раздела 1 из NOMOS

> *Не приносите мне один скриншот, а затем говорите за всю систему.*
>
> *Не делайте ответ одного пользователя ответом всего мирового населения.*
>
> *Не превращайте дубликаты одной и той же учетной записи в независимых людей.*
>
> *Не обновляйте ответ и не отображайте предпочтительное как если бы это было реальным распределением.*
>
> *Не удаляйте отклонения, ошибки и неизвестные значения из знаменателя.*

Не обобщайте один язык на весь мир, одну страну на все страны или один продукт ИИ на все генеративные системы.

Не превращайте один успех в доказательство эффекта вмешательства и одну ошибку в неизменный характер всей системы.

Не топите одну серьезную ошибку в среднем показателе.

Сначала зафиксируйте событие. / Затем измерьте закономерность. / Затем определите целевую популяцию. / Затем покажите зависимость и неопределённость. / И говорите только в той мере, в какой это позволяет доказательная база.

Заключительное предложение главы

Представление сущности в системах искусственного интеллекта не является единственным ответом. Продукты ИИ — это распределение, которое следует измерять по пользователям, странам, языкам, сеансам и времени.

Нормативное ядро

Один выход генеративной системы МОЖЕТ устанавливать, что конкретный ответ произошел при зафиксированных условиях. Один выход НЕ ДОЛЖЕН использоваться для вывода о: - распространенности в популяции, - типичном поведении продукта, - временной стабильности, - согласованности между языками, - согласованности между странами, - согласованности между продуктами, - эффекте вмешательства, - достоверности рекомендаций, - коммерческом воздействии, - соответствие, - или оценке GEO. Все более широкие утверждения ДОЛЖНЫ поддерживаться заранее определённой целевой популяцией, действительным дизайном многократных наблюдений, зафиксированной структурой зависимостей, комплектом доказательств, правилом оценки, заявлением о сфере применения и раскрытием неопределённости. Один выход с высоким уровнем вреда МОЖЕТ вызвать отдельное рассмотрение критического инцидента без того, чтобы считаться доказательством распространенности в популяции.

ГЛАВА 02

Представление — это распределение

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K04; K05; K09

Граница главы

Раздел 1 установил следующее положение:

> *Одиночный ответ не является оценкой GEO. Одиночный ответ — это отдельное наблюдение, произошедшее при зафиксированных условиях.*

Этот раздел отвечает на следующий вопрос:

> *Если один ответ не является баллом GEO, то какой объект мы действительно хотим измерить?*

Ответ:

> *Представление сущности в генеративных системах — это многомерное распределение, формируемое через продукты ИИ, пользователей, страны, языки, условия сессий, промпты и время.*

Эта глава:

- распределение представления,
- основные измерения распределения,
- разница между средним и распределением,
- неравенство подгрупп,
- стабильность,
- проблема охвата и отсутствующих наблюдений,
- риск редких серьезных ошибок,
- концептуальная структура общедоступной карточки результатов

определяет. В этой главе пока не рассматривается:

- сколько участников будет назначено на каждую страну,
- как будет установлен образец выборки,
- как будут проверяться переводы промптов,
- как будут оцениваться атомарные утверждения,
- точная формула итогового NOMOS балла,
- пороговые значения соответствия

Эта глава нормативно не решает эти вопросы; в последующих главах они рассматриваются отдельно. Прежде чем определить формулу оценки, Глава 2 определяет реальность, которую должен представлять балл.

NOMOS Вызов

Рассмотрим два продукта ИИ. Каждый получает одинаковый средний глобальный балл:

90 / 100

Языковые результаты первого продукта следующие:

- Английский: 90
- Турецкий: 91
- Немецкий: 89
- Японский: 90
- Арабский: 90

Языковые результаты второго продукта следующие:

- Английский: 99
- Турецкий: 98
- Немецкий: 97
- Японский: 86
- Арабский: 70

Простое среднее значение обоих продуктов составляет 90. Обладают ли они одинаковым качеством GEO? Нет. Первый продукт обеспечивает схожее представление на разных языках. Второй продукт обеспечивает отличное представление на некоторых языках, но серьезно слабое представление на другом языке. Обычное среднее скрывает эту разницу. Теперь рассмотрим два других продукта. 95% ответов обоих удовлетворяет критериям прохождения. Остальные 5% в первом продукте состоят из:

- незначительных недостатков,
- ограниченных объяснительных недостатков,
- несколько НЕИЗВЕСТНЫХ

результатов. Остальные 5% во втором продукте состоят из:

- неверной информации о здравоохранительных органах,
- несуществующей лицензии,
- неправильной юридической идентичности,
- неподходящих рекомендаций

существует. Стандартная точность для обоих снова составляет 95 процентов. Есть ли у них одинаковый риск представления? Нет. Теперь рассмотрим третью пару. Среднее значение обоих через три волны измерений составляет 90. Первый:

- Волна 1: 90
- Волна 2: 91
- Волна 3: 89

Второй:

- Волна 1: 98
- Волна 2: 90
- Волна 3: 82

Имеют ли они одинаковую стабильность во времени? Нет. В этом случае проблема заключается не только в следующем: «Каково среднее?» Нам также необходимо знать следующее: Как распределены результаты? Кто хорошо представлен, кто плохо представлен?

Какие языки и страны отстают? Насколько серьезны ошибки? Насколько результат меняется со временем? Какие области никогда не измерялись? С какими весами было создано среднее? Сколько различных реальностей скрывается за одним средним? Первый вывод этого раздела таков:

> *Одинаковое среднее не является одинаковым представлением.*

Второе положение гласит:

> *Оценка GEO может быть сводкой распределения; она не может заменить распределение.*

1. ЦЕЛЬ ГЛАВЫ

В этом разделе определяется, как следует понимать представление сущностей в генеративных системах до того, как оно будет сведено к единому числу. Его основные цели:

- Показать, что представление не является вневременной и единственной «точкой зрения ИИ»
- Определить условия, от которых зависит результат представления
- Объяснить, что бинарная система true–false сама по себе недостаточна
- Разделить среднюю точность и неравенство между подгруппами
- Сделать видимым высокий риск ошибок, скрытый за высокими средними значениями
- Показать, что стабильность и точность — это не одно и то же
- Предотвратить заполнение неизмеренных областей положительными предположениями внутри распределения
- Подготовить структуру будущего показателя NOMOS, который является единичным, но не обречен быть единичным

В конце этого раздела читатель должен быть в состоянии ответить на следующий вопрос:

> *Почему результат GEO сущности не ограничивается только средней точностью?*

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

> *Представление сущности в генеративных системах — это условное распределение различных ответов и результатов представления, которые могут возникать при определённых условиях.*

Это распределение:

- это не абстрактная «мысль» одной модели,
- это не общее мнение всех систем искусственного интеллекта,
- это не только частота повторений на сайте,
- Дело не только в частоте упоминания названия бренда,
- это не просто средняя точность,

Это не только показатель рекомендаций. Распределение представления в рамках определенного протокола показывает следующее:

> *Какие пользователи видят какой вид представления на каком продукте ИИ и пользовательский интерфейс, на каком языке, в какой стране, с каким запросом, при каких условиях сеанса и в какое время.*

Следовательно, результат распределения всегда зависит от следующих вопросов: Какая сущность? Какой продукт ИИ? Какой пользовательский интерфейс? Какая целевая аудитория? Какая страна? Какой язык и локализация? Какой запрос? Какие условия сеанса? Какая волна измерений? Какой Truth Pack? Какое правило оценки? Когда любое из этих полей изменяется, распределение также может измениться.

3. ЧТО ОЗНАЧАЕТ СЛОВО «РАСПРЕДЕЛЕНИЕ»?

В этой книге слово "распределение" не относится к скрытому или метафизическому взгляду, который предполагается существующим внутри системы искусственного интеллекта. Оно не утверждает: NOMOS: "Это реальное распределение компании в разуме модели." Здесь под распределением подразумевается операционное и наблюдаемое распределение.

> *Это схема различных результатов представления, которые могут быть получены в рамках определенного протокола в целевой популяции и условиях измерения.*

Следует различать два распределения.

3.1. Распределение целевых представлений

Распределение, которое измерение пытается предсказать, но не может полностью наблюдать напрямую. Оно представлено следующим образом:

P_E(Y | A, C, L, S, P, T, U)

Здесь:

- E: проверяемая сущность
- Y: результат представления
- A: продукт ИИ и пользовательский интерфейс
- C: страна или географические условия
- L: язык и локаль
- S: условия сессии и персонализации
- P: запрос
- T: время или волна измерения
- U: целевая вселенная пользователей

Выражение задает вопрос: «В определенной целевой вселенной пользователей, при указанных условиях, с какими вероятностями происходят результаты представления?» Не все элементы целевого распределения могут быть измерены. Они оцениваются через выборку.

3.2. Наблюдаемое эмпирическое распределение

Это распределение действительных наблюдений, полученных в реальном или синтетическом измерении. Его представление:

P̂_E

может быть использовано. Эмпирическое распределение само по себе не является целевым распределением. Это оценка, полученная через его выборку, взвешивание и дизайн измерений. Величина различий может зависеть от:

- Дизайн выборки
- Выбор участников

- Охват стран и языков
- Доля некорректных наблюдений
- Отсутствие ответов
- Ошибка арбитра
- Измерительная волна
- Выбор промптов
- Условия продукта и интерфейса
- Взвешивание
- Кластеризация и зависимость

Следовательно, эмпирическое распределение не может быть опубликовано как «абсолютная и неизменная истина искусственного интеллекта». Правильное выражение должно быть: «Оценка, полученная для указанной целевой популяции и условий измерений».

4. ПРОСТРАНСТВО ПРЕДСТАВЛЕНИЯ

Представление сущности не происходит в одном измерении. Каждое наблюдение существует в многомерном пространстве условий. Базовое представление:

```
X_i= (E_i, A_i, U_i, C_i, L_i, S_i, P_i, T_i)
```

Здесь:

- Ei: сущность
- ИИ: продукт ИИ и пользовательский интерфейс
- Ui: пользователь или единица выборки
- Ci: страна и географический слой
- Li: язык и локаль
- Si: сеанс, план и условие персонализации
- Pi: запрос или семейство запросов
- Ti: время и волна измерений

Сгенерированный ответ:

```
R_i
```

можно показать. Представительное состояние, полученное в результате оценки ответа, следующее:

```
Y_i= g(R_i, V_E, J)
```

и может рассматриваться как. Здесь:

- VE: использован проверенный пакет Truth Pack для объекта
- J: правила оценки и вынесения решений
- g: процесс оценки, который преобразует ответ в представительные результаты

Значение этого представления следующее: один и тот же ответ не может быть оценен самостоятельно без указания, по какой сущности, какому Пакету Истины и какому правилу оценки он был проверен.

5. ПРЕДСТАВИТЕЛЬНЫЙ РЕЗУЛЬТАТ НЕ ЯВЛЯЕТСЯ ОДНИМ ЗНАЧЕНИЕМ

Только целостность ответа:

- не может быть вынуждено к категориям

  истинно,
- неверна

Присвоение метки часто недостаточно. Следующие ситуации могут происходить одновременно в одном и том же ответе:

- Правильная идентичность
- Правильная основная категория
- Неполное покрытие услуг
- Устаревшее количество сотрудников
- Неподтвержденное утверждение о руководстве
- Правильная цитата
- Неподходящая рекомендация
- Некритичный дефицит границ

Следовательно, результат представления можно рассматривать как вектор результата. Кандидат на представление:

```
Y_i= (y_i^id, y_i^cat, y_i^fact, y_i^scope, y_i^time, y_i^evidence, y_i^recommendation, y_i^harm)
```

Здесь:

- yiid: точность идентичности
- yicat: точность категории и роли
- yifact: точность фактических данных
- yiscope: сохранение объема и границ
- yitime: временная достоверность
- yievidence: согласованность утверждения и доказательств
- yirecommendation: рекомендации и пригодность для пользователя
- yiharm: класс вреда ошибки или неверного руководства

Точные правила оценки этих полей будут определены в последующих разделах. Основной принцип здесь:

> *Представительный ответ должен оцениваться не только по тому, содержит ли он правильную информацию, но и по тому, о ком, о чем, в каких границах, с какими доказательствами и для кого он представлен надлежащим образом.*

6. ОСНОВНЫЕ СОСТОЯНИЯ ПРЕДСТАВЛЕНИЯ

Статус каждого отдельного утверждения и окончательного ответа может различаться. В этом разделе концептуально выделяются следующие основные состояния.

6.1. ПОДТВЕРЖДЕНО

Утверждение подтверждается достаточными доказательствами в пределах определённого объёма и времени.

6.2. ОПРОВЕРГНУТО

Утверждение противоречит достоверным и релевантным записям.

6.3. НЕПОДТВЕРЖДЕНО

Для утверждения недостаточно подтверждающих материалов. Не доказано, что утверждение обязательно ложное. Однако записей недостаточно, чтобы принять его за истину.

6.4. УСТАРЕЛО — БОЛЬШЕ НЕ АКТУАЛЬНО

Утверждение могло быть верным в прошлом. На дату измерения оно больше не является действительным или его текущую точность невозможно проверить.

6.5. ЧАСТИЧНО ПОДТВЕРЖДЕНО — ЧАСТИЧНО ПОДТВЕРЖДЕНО

Поддерживается только определенный объем, население, страна, продукт или время утверждения. Ответ мог превысить лимит поддержки.

6.6. ВНЕ СФЕРЫ — ВНЕ СФЕРЫ

Утверждение не связано с проверяемым объектом или Truth Pack. Оно могло исходить от другой организации, продукта, страны или услуги.

6.7. СМЕШЕНИЕ СУБЪЕКТОВ — СЛИЯНИЕ СУБЪЕКТОВ

Система:

- две разные компании,
- бренд и юридическое лицо,
- человек и учреждение,
- старое и новое юридическое лицо,
- перепутала разные организации с одинаковыми названиями

друг с другом. Эта ошибка не является простым фактическим недочетом. Это искажение объекта представления.

6.8. ВАЖНОЕ ПРОПУЩЕННОЕ — ОТСУТСТВУЮЩАЯ ИНФОРМАЦИЯ

Предложения в ответе могут быть правильными. Однако важная информация, которая могла бы повлиять на решение пользователя, была опущена. Пример: услуга предоставляется только в определённых странах. Лицензия ограничена конкретной юрисдикцией. Цена указана без учета налога. Результат основан на данных первой стороны. Продукт не подходит для определённых пользователей. Опущение может привести к неправильным выводам, даже если нет напрямую неверного предложения.

6.9. ОТКАЗ — ОТКЛОНЕНИЕ

Система отказалась отвечать. Отказ:

- соответствующий,
- ненужный,
- несогласованный,

- основанный на безопасности,
- технический

может возникнуть. Ret не считается автоматически неудавшимся или успешным. Он классифицируется отдельно.

6.10. НЕТ РЕЗУЛЬТАТА

Система не смогла предоставить полезный ответ по техническим или операционным причинам. Это не то же самое, что Ret.

6.11. НЕИЗВЕСТНО

Доступные записи недостаточны для надежной классификации. НЕИЗВЕСТНО не может быть преобразовано в положительный результат.

7. ОСНОВНЫЕ ХАРАКТЕРИСТИКИ РАСПРЕДЕЛЕНИЯ

Распределение представления не может быть объяснено только средним значением. Как минимум, следует оценить следующие характеристики.

7.1. Центральная тенденция

Суммирует общую точность или уровень перехода ответов. Примеры показателей:

- Среднее
- Взвешенное среднее
- Медиана
- Доля успешных результатов

Центральная тенденция важна. Однако она показывает только центр распределения.

7.2. Разброс и изменчивость

Указывает, насколько результаты разбросаны вокруг центра. У двух систем с одинаковым средним одно может быть последовательным среди пользователей, в то время как другое может быть очень изменчивым. Изменчивость:

- пользователей,
- стран,
- языков,
- сессий,
- волны измерений

также можно изучать среди них.

7.3. Форма распределения

Сосредоточены ли результаты вокруг одной центральной точки? Или они разделены на два и более кластеров? Например:

- Если половина пользователей видит очень точные,
- а половина видит очень неточные

ответы, среднее может быть средним. Однако типичный опыт пользователя не является «средней точностью». Это два отдельных поляризованных опыта.

7.4. Различия между подгруппами

Видят ли определенные языковые, страновые или пользовательские группы систематически разные результаты? Глобальное среднее может скрывать эту разницу.

7.5. Риск в хвосте

Существуют ли редкие, но серьезные ошибки? Редкие события могут оказывать небольшое влияние на среднее значение. Их влияние на вред для человека может быть значительным.

7.6. Покрытие

Какую часть целевой аудитории действительно измерили? Высокий результат может сочетаться с низким покрытием.

7.7. Пропавшая и неизвестная масса

Какая доля ответов:

НЕИЗВЕСТНО,

НЕ ТЕСТИРОВАЛАСЬ,

отказ, технический ответ отсутствует или недействительная запись?

7.8. Временная стабильность

В какой степени распределение сохраняется на разных волнах измерений?

8. КАНДИДАТ NOMOS ПОДПИСЬ ПРЕДСТАВИТЕЛЬНОГО РАСПРЕДЕЛЕНИЯ

КАНДИДАТСКАЯ КОНЦЕПЦИЯ

Рекомендуется, чтобы результат GEO сначала определялся следующей подписью распределения, а не отдельным числом:

```
RDS(E) = (μ, κ, σ, Δ, τ, ρ, υ)
```

Это представление называется Подпись Распределения Представлений, а именно:

Подпись Распределения Представлений

Компоненты:

μ — Общий уровень представления

Общая точность или прогноз перехода при заданных весах и области.

κ — Покрытие

Степень, в которой целевой пользователь, страна, язык, ИИ-продукт и временной интервал достаточно измерены.

σ — Вариабельность

Рассеяние среди пользователей и экспериментальных ячеек.

Δ — Неравенство подгрупп

Существенная разница между языками, странами, продуктами или другими значимыми группами.

'tau — Риск критических и хвостовых ошибок

Частота и характер редких серьёзных ошибок.

ρ — Временная стабильность

Устойчивость распределения по волнам.

u — Неизвестная и неразрешенная масса

НЕИЗВЕСТНО — это доля распределения, которая остается неразрешенной или недостаточно оцененной. Эта сигнатура пока не является окончательной формулой оценки. Следующие элементы требуют дальнейшей разработки:

- определённое определение,
- оценка,
- вес,
- публичная презентация

Они будут завершены в ходе пилотных исследований и в последующих главах. Основной принцип заключается в следующем:

> *Результат GEO должен давать ответ не только на вопрос «насколько точно?», но и «для кого, насколько последовательно, насколько справедливо, насколько полно и насколько рискованно?»*

9. СРЕДНЕЕ ОТБЕЛИВАНИЕ

Если учреждение демонстрирует только высокий глобальный средний показатель и скрывает серьёзные недостатки распределения, это называется:

Среднее отбеливание

Среднее значение может быть рассчитано следующими способами:

- Сокрытие самого слабого языка
- Исключение маленьких стран из отчета
- Смешивание критических ошибок с общей ошибкой
- Удаление результатов НЕИЗВЕСТНОГО из знаменателя
- Учет только хорошо работающих ИИ-продуктов
- Исключение слабой волны измерений
- Сделать невидимой популяцию, находящуюся вне сферы охвата
- Изменение весов после получения результатов

Среднее может быть вычислено математически правильно. Однако представленная им информация для общественности может вводить в заблуждение. Поэтому:

> *Математически правильное вычисление среднего не означает, что оно достаточно для принятия решений.*

10. РАЗНЫЕ РЕАЛЬНОСТИ ОДНОГО И ТОГО ЖЕ СРЕДНЕГО

Среднее значение может быть 90 во всех следующих синтетических распределениях.

10.1. Однородное распределение

Все основные подгруппы находятся в диапазоне 88–92. Это распределение относительно сбалансировано.

10.2. Поляризованное распределение

Половина пользователей видит результаты, близкие к 100, в то время как другая половина видит результаты, близкие к 80. Среднее значение — 90. Однако пользовательский опыт неодинаков.

10.3. Лингвистически фрагментированное распределение

Результаты очень хорошие для языков с высокой численностью населения или большими ресурсами, но значительно слабые для языков с ограниченными ресурсами. Глобальное среднее может выглядеть хорошо. Языковая справедливость слабая.

10.4. Распределение по странам с фрагментацией

Даже внутри одного языка результаты могут различаться из-за разных стран, регионов, доступа к продукту или пользовательской базы.

10.5. Временно нестабильное распределение

Очень высокие результаты происходят в одной волне, очень низкие результаты — в другой волне. Даже если среднее значение высоко, трудно установить надежные ожидания пользователей.

10.6. Распределение с хвостовым риском

Большинство ответов верны. Однако небольшая часть содержит серьезные ошибки, связанные со здоровьем, правом, лицензированием или безопасностью.

10.7. Распределение высокой неизвестной массы

Большинство оцениваемых ответов верны. Однако значительная часть целевой аудитории или ответов:

- подверглась тестированию,
- не тестировалась,

НЕИЗВЕСТНО

или результаты неубедительны. Публикация высокого показателя, основанного только на оцениваемых ответах, может скрыть масштаб.

11. НЕРАВЕНСТВО ПОДГРУПП

Глобальный показатель может не отражать опыт внутри подгрупп. Ключевые подгруппы могут включать:

- Продукт ИИ и пользовательский интерфейс
- Страна
- Язык
- Локаль
- План

- Веб или мобильное
- Контролируемая или естественная панель
- Временная волна
- Намерение пользователя
- Семейство устройств

Не все эти группы могут иметь одинаковую важность в каждом аудите. Основные группы должны быть определены после утверждения области.

11.1. Язык — это не то же самое, что страна

Туркоязычные:

- Пользователь в Турции,
- Пользователь в Германии,
- Пользователь в другой стране

Они могут использовать один и тот же язык. Однако:

- доступ к продукту,
- цена,
- ресурсы,
- локальный интерфейс,
- юридический контекст,
- персонализацию

могут различаться. Следовательно, все различия не могут объясняться только языком или страной.

11.2. Интерсекционные клетки

Некоторые материальные различия становятся видимыми только при одновременном рассмотрении нескольких областей. Пример: Германия × Турция × Мобильный × Бесплатный тариф × Панель пользователей в естественных условиях сама по себе может быть экспериментальной ячейкой. Однако по мере увеличения количества ячеек размер выборки уменьшается. Поэтому окончательный результат нельзя публиковать для каждой маленькой ячейки. Дизайн измерения: должен сохранять значимые подгруппы и не должен создавать статистически ложную точность.

11.3. База подгруппы

КАНДИДАТСКАЯ КОНЦЕПЦИЯ

В дополнение к глобальному среднему значению рекомендуется отдельно сообщать результат самой слабой адекватно исследованной подгруппы или группы нижнего 10%. Это можно назвать:

Основа представления

или:

Язык/Страна Пол

Однако, если «наименьший балл» получается из маленькой и нестабильной ячейки, это может дать неправильные результаты. Поэтому базовая метрика должна быть определена вместе с:

- минимальный образец,

- доверительный интервал
- качество ячейки

Точное правило будет определено в следующих разделах.

12. ЭТО ЧАСТЬ РАСПРЕДЕЛЕНИЯ ОБЛАСТИ ОХВАТА

Распределение состоит не только из наблюдаемых ответов. То, что никогда не наблюдалось, также определяет значение результата. Например, были протестированы английский и турецкий языки; японский и арабский — нет. Веб-интерфейс был протестирован; мобильный интерфейс и платная версия — нет. Сильные результаты для английского и турецкого языков не оправдывают утверждение: «Представление хорошо во всех языках и на всех пользовательских интерфейсах». Обоснованным выводом является: «Измерения проводились на указанных веб-платформах для английского и турецкого языков. Другие языки и мобильные платформы не тестировались.»

12.1. Карта покрытия

Каждое распределение GEO должно содержать следующую карту в той мере, в какой это актуально:

- Целевая вселенная
- Выборочная совокупность
- Покрытые страны
- Непокрытые страны
- Покрытые языки
- Непокрытые языки
- Мониторируемые ИИ-продукты
- Немониторируемые продукты
- Мониторируемые пользовательские интерфейсы
- Немониторируемые интерфейсы
- Волны измерений
- Семейства промптов вне области

12.2. Уровень покрытия

КАНДИДАТСКАЯ КОНЦЕПЦИЯ

Можно разработать метрику покрытия, чтобы показать, какая часть целевой вселенной представлена с достаточным уровнем измерения. Однако покрытие:

- только количество стран,
- только количество языков,
- только доля населения

не может быть измерено. Например, может быть охвачено 150 стран. Однако 95% участников могли прийти только из пяти стран. Или может быть охвачено 90% населения, но малые языки могут быть полностью исключены. Поэтому охват населения следует показывать отдельно от охвата стран/языков наблюдателями.

13. ПРОПАВШИЕ НАБЛЮДЕНИЯ И НЕОТВЕТЫ

Отсутствующие данные — это не только техническая проблема. Их распределение может быть систематически искажено. Причины отсутствия наблюдения могут включать:

- Неудовлетворительное выполнение задачи участником
- Продукт недоступен в стране
- Ограничения вычислений
- Техническая ошибка
- Отклонение системой
- Неверный снимок экрана
- Изменение запроса участником
- Продукт не работает на определённом языке
- Судьи, неспособные принимать надежные решения

Каждая из этих причин требует отдельного статуса.

13.1. Отсутствующие данные могут быть не случайными

Например, если больше заданий не могут быть выполнены в странах с низким качеством соединения, использование только успешных записей может сместить глобальный результат в пользу состоятельных пользователей с сильным соединением. Если уровень НЕИЗВЕСТНО повышается из-за того, что ответы на определенном языке трудно оценить, оценка, рассчитанная только из известных ответов, может скрывать языковое неравенство. Поэтому:

> *Отчет о пропавших данных и распределении причин отсутствующих данных должен быть частью отчета.*

13.2. Знаменатель — целостность знаменателя

Отчет должен ясно ответить на вопрос: «90 процентов от чего именно составляет 90 процентов?» Возможные знаменатели:

- Назначенные участники
- Инициаторы задач
- Исполнители заданий
- Записи, соответствующие протоколу
- Действительные системные результаты
- Ответы с решениями арбитра
- Ячейка конкретного языка или страны

Знаменатель не должен изменяться после того, как результат виден. Например:

- Назначенные задания 1,000,
- Выполненные задания 900,
- Действительные задания 800,
- Пройденные задания 720

Таким образом:

```
720/800=90%
```

Это коэффициент успеха среди действительных записей. Наблюдаемый коэффициент успеха по всей назначенной выборке вместо этого:

```
720/1000=72%
```

отвечает на другой вопрос. Оба показателя можно использовать. Их нельзя представлять в одинаковом смысле.

14. ТОЧНОСТЬ И ПОСЛЕДОВАТЕЛЬНОСТЬ — ЭТО НЕ ОДНО И ТО ЖЕ

Распределение может иметь высокую точность. Однако оно может быть непоследовательным во времени или между пользователями. Другое распределение может иметь низкую точность. Однако оно может быть систематически неправильным. Поэтому следует различать четыре ситуации.

Точность	Последовательность	Комментарий
Высокая	Высокая	Кандидат на последовательное и точное представление
Высокая	Низкая	В основном точный, но ненадежный опыт
Низкая	Высокая	Систематическое и последовательное искажение
Низкая	Низкая	И неправильное, и нестабильное представление

Быть «последовательным» не всегда положительно. Система может быть довольно последовательной, если она дает одну и ту же ошибку всем пользователям.

14.1. Согласованность между пользователями

Видят ли разные пользователи одинаковое представление материала в одной и той же ячейке?

14.2. Согласованность между сессиями

Одинаковый или сопоставимый пользователь видит аналогичные результаты в новых сессиях?

14.3. Межволновая согласованность

Производит ли один и тот же протокол схожее распределение в разные дни или периоды?

14.4. Согласованность после обновления продукта

Когда в продукте ИИ происходят существенные изменения, продолжается ли прежнее распределение?

15. ПРОИЗВОДСТВО ОДНОГО И ТОГО ЖЕ ПРЕДЛОЖЕНИЯ НЕ ЯВЛЯЕТСЯ СТАБИЛЬНОСТЬЮ

Стабильность: это не означает «давать каждому пользователю тот же самый ответ слово в слово». Продукт ИИ может использовать разные выражения, порядок и примеры. Пока сохраняется материальный смысл, это могут быть приемлемые варианты. Важное различие заключается в следующем:

Формальная изменчивость

Выбор слов Порядок предложений Длина Тон Пример Использование списка или абзаца Если сохраняется материальная истина, это может не быть ошибкой.

Семантическая изменчивость

Смена идентичности компании Дифференциация основной категории Расширение или сокращение области деятельности Смена лицензии или компетенции Изменение цены и географии Рекомендация одному пользователю, но не аналогичному пользователю без объяснения Если материальное представление меняется, может возникнуть проблема стабильности. Поэтому предпочтительной концепцией является:

Условная семантическая стабильность

Так должно быть. Ожидается сохранение фундаментальной реальности при эквивалентных материальных условиях. Изменение рекомендации из-за разных потребностей пользователя может быть соответствующим контекстным поведением.

16. ЧУВСТВИТЕЛЬНОСТЬ К КОНТЕКСТУ И ПРОИЗВОЛЬНАЯ ВАРИАТИВНОСТЬ

Система не всегда должна давать одинаковые рекомендации всем пользователям. Например:

- Пользователи с низким бюджетом
- Только пользователи, ищущие услуги в конкретной стране
- Пользователи, требующие иной лицензии
- Пользователи, ищущие индивидуальные, а не корпоративные услуги

могут получать различные результаты соответствия для одной и той же сущности. Эта разница:

- может быть обусловлена состоянием пользователя,
- объемом услуги,
- или фактическим лимитом

и может быть чувствительной к контексту и точной. Однако если результат меняется без причины при равных условиях пользователя, это может быть произвольной вариабельностью. Поэтому NOMOS не хочет: «Одинаковый ответ для всех пользователей». Оно хочет:

> *Ответ, который является последовательным и объяснимым при эквивалентных материальных условиях и чувствительным к фактическим пределам при различных условиях.*

17. ВРЕМЯ ВНУТРИ РАСПРЕДЕЛЕНИЯ ПРЕДСТАВЛЕНИЯ

Распределение представления не вневременно. Его можно рассматривать как PE,t. Распределение одной и той же сущности:

- обновление продукта,
- изменение ресурсов,
- изменение компании,
- новостное событие,
- новая языковая версия,
- техническое изменение доступа

может расходиться впоследствии.

17.1. Внезапное изменение

После обновления продукта или ресурса распределение меняется быстро за короткое время.

17.2. Постепенное изменение

Из-за новых ресурсов и поведения пользователей результат постепенно изменяется со временем.

17.3. Временные колебания

Во время конкретного события или новостного периода распределение временно отличается.

17.4. Изменение после вмешательства

После вмешательства GEO распределение может измениться. Однако для утверждения, что изменение произошло из-за вмешательства, также требуется базовая линия и проект причинно-следственной связи.

17.5. Сглаживание средним за время

Для продукта:

- в первой волне 98,
- во второй волне 90,
- в третьей волне 82

Если взять среднее, получится 90. Однако публикация 90 скрывает нисходящий тренд. Следовательно:

> *Временное среднее не может заменить результаты по волнам.*

18. РИСК ХВОСТА И РАСПРЕДЕЛЕНИЕ СИЛЬНЫХ ОШИБОК

Большая часть распределения может быть правильной. Однако следует также оценивать ошибки с низкой вероятностью и большой величиной. Примеры:

- Лицензия на медицинскую деятельность, которая на самом деле не существует
- Неправильное руководство по чрезвычайным ситуациям
- Неправильные юридические полномочия
- Фальшивая гарантия инвестиций
- Неправильная идентичность компании
- Неподходящие советы для детей или уязвимых групп
- Повторение манипулятивной инструкции, скрытой от людей

В предыдущей архитектуре:

- Критический,
- Серьезный,
- Умеренный,
- Консультативный

были разделены на четыре уровня важности, и предполагалось, что критические находки должны автоматически вызывать отказ. В этой книге эта система связывается с распределением населения.

18.1. Разделение частоты и серьезности

Ошибка:

- частая, но с низким воздействием,
- редкая, но серьезная

может возникнуть. Однако только частота не может одинаково оценить оба случая.

18.2. Представление кандидата на основе риска с учетом серьезности

КАНДИДАТСКАЯ КОНЦЕПЦИЯ

Индикатор риска может быть разработан с использованием коэффициентов серьезности для классов ошибок:

```
tau= Σ_{k=1}^K λ_k p_k
```

Здесь:

- pk: прогнозируемая частота данного типа ошибки
- λk: коэффициент тяжести типа ошибки

Однако коэффициенты не могут определяться произвольно. Они должны основываться на:

- Предсказуемый вред для человека
- Целостность представления
- Влияние на решение пользователя
- Обратимость
- Намерение и манипуляция
- Зона повышенного риска

Точный метод будет определён в Разделе 15.

18.3. Критический этап

Высокий общий средний показатель не может автоматически исключать проверенного кандидата с критической ошибкой. Правильная публичная запись должна показывать оба результата вместе: «Оценка общего представления высока.» и: «Указанное критическое событие или события также влияют на решение о соответствии.»

19. ОТКАЗ И НЕУСПЕХ В РАСПРЕДЕЛЕНИИ

Система, которая не отвечает, также является частью пользовательского опыта. Однако не все отказы одинаковы.

19.1. Соответствующий отказ

Система может проявлять осторожность, когда доказательств недостаточно или когда она не может предоставить безопасный ответ. В некоторых контекстах это может быть правильным поведением.

19.2. Ненужный отказ

Система может непоследовательно избегать ответа на общедоступную и низкорискованную информацию.

19.3. Техническое отсутствие ответа

Система может не дать ответа из-за технической ошибки.

19.4. Отказ из-за выхода за рамки

Запрос может быть неправильным или выходить за рамки протокола задачи.

19.5. Сохранение отказов в распределениях

Отказы:

- успех,
- неудача,

НЕИЗВЕСТНО

не следует заставлять их попадать в одну категорию. Следует оценивать их функцию и обоснованность. Исключение отклонений из знаменателя может дать высокий уровень точности только среди успешных ответов и исказить реальный пользовательский опыт.

20. КАК МОЖНО СРАВНИВАТЬ РАСПРЕДЕЛЕНИЯ?

Просто сравнивать численные результаты недостаточно для сравнения распределения двух ИИ-продуктов. Чтобы сравнение было справедливым, следующие области должны быть максимально равными или явно смоделированы:

- Проверяемый объект
- Целевая популяция
- Версия промпта
- Языки
- Страны
- Измерительная волна
- Условия сессии
- пользовательский интерфейс
- Распределение планов
- Правило вынесения решения
- Пакет реальности
- Правила достоверности
- Весы

Два показателя, полученные с существенно различными протоколами, нельзя использовать напрямую для ранжирования.

20.1. Сравнение на общей основе

Все сравниваемые ИИ-продукты:

- официально доступны,
- поддерживают один и тот же язык и условия использования

могут быть сравнены в общем пространстве. Это обеспечивает более чистое сравнение между системами.

20.2. Распределение естественного доступа

Каждый ИИ-продукт может быть измерен в собственной реальной вселенной пользовательского доступа. Этот результат показывает фактическое глобальное влияние продукта. Однако, поскольку целевые аудитории могут отличаться, прямое ранжирование по производительности требует осторожности.

20.3. Разные интерфейсы одного провайдера

Веб-, мобильные, корпоративные продукты или разные планы, даже если они принадлежат одному поставщику, могут создавать отдельные распределения. Если их необходимо объединить в один результат, необходимо объяснить метод комбинации и веса.

21. ВЕС ИЗМЕНЯЕТ ЗНАЧЕНИЕ РАСПРЕДЕЛЕНИЯ

Глобальное среднее зависит от веса, с которым объединяются наблюдения. Возможные веса:

- Равные по странам
- По населению стран
- Соответствующая популяция пользователей ИИ
- Пользователи по языку
- Фактическое воздействие продукта
- Равные продукты ИИ
- Доля рынка
- Целевой рынок клиентов

Каждый метод отвечает на разный вопрос.

21.1. Равный вес по странам

Сан-Марино и Индия могут получать одинаковый вес. Этот подход может решить вопрос «среднего опыта страны». Он не отражает средний пользовательский опыт во всем мировом населении.

21.2. Вес населения

Крупные страны имеют большее влияние. Этот подход может приблизиться к опыту мирового населения. Он может сделать маленькие страны и языки невидимыми.

21.3. Вес языковой справедливости

Это также может обеспечить видимость языков с ограниченными ресурсами. Это не то же самое, что глобальный показатель населения.

21.4. Вес экспозиции продукта

Можно учитывать взвешивание продуктов ИИ в зависимости от фактической пользовательской экспозиции. Это требует надежных и независимых данных о использовании.

21.5. Принцип нормативного взвешивания

> *Веса должны быть определены до того, как будут видны результаты, публично раскрыты, и следует указать вопрос, на который они отвечают.*

Запрещается выбирать систему взвешивания, которая будет увеличивать положительный результат после того, как результаты стали известны. Наряду с единой глобальной оценкой могут также публиковаться альтернативные значимые взвешенные результаты.

22. РАСПРЕДЕЛЕНИЕ ОЦЕНОК — С ПОТЕРЯМИ

Возможно получение одного числа из распределения. Однако одно число не несет всю информацию, содержащуюся в распределении. Поэтому оценка:

Является сжатым с потерями представлением распределения

должна приниматься. Оценка может терять следующую информацию:

- Слабейший язык
- Слабейшая страна
- Тип критической ошибки
- Неизвестная масса
- Колебания во времени
- Поляризация
- Различие продукта и плана
- Неподходящая популяция
- Распределение отклонений
- Неопределённость арбитра

Оценка полезна, потому что:

- облегчает сравнение,
- упрощает публичное представление,
- позволяет отслеживать изменения,

поддерживает управленческое решение. Однако простота оценки не может превращаться в устранение распределения.

23. МИНИМАЛЬНАЯ КАРТА ПУБЛИЧНЫХ РЕЗУЛЬТАТОВ

ТРЕБОВАНИЕ КАНДИДАТА — ОБЯЗАННОСТЬ КАНДИДАТА

Если глобальная оценка NOMOS будет представлена публике, рекомендуется показать хотя бы следующие результаты вместе с ней:

Поле	Описание
Основная оценка представления	Общий результат с учётом определённого взвешивания
Доверительный или диапазон неопределённости	Чувствительность прогноза
Область	Страна, язык, продукт, пользовательский интерфейс и время
Наиболее слабая достаточная подгруппа	Делает материальное неравенство видимым
Критический уровень ошибок	Очередь серьезных ошибок
Временная стабильность	Вариация между волнами

Поле	Описание
Уровень НЕИЗВЕСТНЫХ и неопределенных результатов	Неизвестная масса
Поля НЕ ТЕСТИРОВАЛИСЬ	Неизмеренная совокупность
Дата измерения	Временное ограничение результата
Версия протокола	По какому правилу был получен результат

Основной показатель не должен использоваться как самостоятельный рекламный ярлык. Каноническая запись должна быть связана с этой карточкой результатов.

24. СИНТЕТИЧЕСКОЕ APPLE.COM ОТОБРАЖЕНИЕ РАСПРЕДЕЛЕНИЯ

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Ниже указанные продукты ИИ, пользователи, языки, числа и результаты полностью вымышлены. Они не отражают реальные показатели Apple Inc. или какого-либо настоящего продукта ИИ. Рассмотрим два синтетических продукта ИИ:

- Продукт A
- Продукт B

Для каждого продукта:

- пять языков,
- 200 действительных синтетических наблюдений на язык,
- всего предполагается 1,000 наблюдений

Это равномерное распределение по языкам предназначено только для демонстрации метода. Оно не отражает реальное распределение GEO-1000 в популяции.

24.1. Процент прохождения по языкам

Язык	Продукт A	Продукт B
Английский	90%	99%
Турецкий	91%	98%
Немецкий	89%	97%
Японский	90%	86%
Арабский	90%	70%
Простое среднее	90%	90%

Среднее значение двух продуктов одинаково. Однако их распределения различаются.

Продукт A

Результаты по языкам близки друг к другу. Разница между самым высоким и самым низким языком составляет 2 баллов. Минимальный уровень языка составляет 89%.

Продукт B

Дает очень высокие результаты по трем языкам. Наблюдается снижение по японскому и особенно по арабскому. Разница между самым высоким и самым низким языком составляет 29

баллов. Минимальный уровень языка составляет 70%. Примечание: фраза “Оба набрали 90” скрывает существенную разницу.

24.2. Распределение статусов ответа

Статус ответа	Продукт A	Продукт B
Соответствие критерию перехода	900	900
Частично поддерживается	60	20
Неподтвержденное утверждение о материале	20	20
Устарело	10	20
Кандидат на критическую ошибку	1	14
НЕИЗВЕСТНО / нет результата	9	26
Итого	1,000	1,000

Проходной балл двух продуктов снова составляет 90%. Однако: в Продукте B количество кандидатов с критическими ошибками в 14 раз больше. В Продукте B масса неизвестных и неопределённых данных выше. Большая часть сбоев в Продукте A связана с частичной нехваткой поддержки. Задолженность по ошибкам Продукта B выше. Один и тот же проходной балл не обязательно приводит к одинаковым результатам пригодности.

24.3. Волны времени

Измерительная волна	Продукт A	Продукт B
Волна 1	90%	98%
Волна 2	91%	90%
Волна 3	89%	82%
Среднее	90%	90%

Продукт A относительно стабилен. Продукт B демонстрирует заметное снижение. Публикация только среднего значения трёх волн скрывает ухудшение в продукте B.

24.4. Синтетический результат

Возможно опубликовать одинаковый заголовочный балл для продукта A и продукта B:

90

Но NOMOS карточки результатов не могут быть одинаковыми.

Продукт A — Синтетическая подпись распределения

Общий результат: высокий
Языковое неравенство: низкое
Хвост критических ошибок: низкий, но не нулевой
Масса неизвестного: низкая
Стабильность во времени: высокая

Продукт B — Синтетическая сигнатура распределения

Общий результат: высокий
Языковое неравенство: высокое
Хвост критических ошибок: значительный
Масса неизвестного: выше

Стабильность во времени: низкая и снижающаяся

Таким образом, синтетический вывод Раздела 2 следующий:

> *Одинаковый средний балл у Продукта A и Продукта B не означает, что у них одинаковое качество представления.*

25. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH02-N01

Результат GEO нельзя определить только по одному среднему показателю или одной пропускной способности.

CH02-N02

Каждый результат распределения:

- аудированная сущность,
- целевая популяция,
- Продукт ИИ и пользовательский интерфейс,
- страна и язык,
- Запрос,
- условие проживания,
- по времени

должно быть подключено.

CH02-N03

Глобальный средний показатель сам по себе не может компенсировать неравенство подгрупп по материальному признаку.

CH02-N04

Критические и другие ошибки высокой степени тяжести следует сообщать отдельно от общей частоты ошибок.

CH02-N05

Статусы UNKNOWN, no-result, rejected, invalid и NOT TESTED следует различать друг от друга.

CH02-N06

Невозможность измерения страны, языка, продукта, пользовательский интерфейс или периода времени не может быть показана как часть положительного распределения.

CH02-N07

Вес и правила агрегации должны быть определены до просмотра результатов.

CH02-N08

Изменения веса после получения результата должны создавать новую версию анализа, а предыдущие результаты должны быть сохранены.

CH02-N09

Распределения, основанные на принципиально разных целевых популяциях, спросе, продукте, времени или критериях оценки, нельзя сравнивать напрямую без объяснения.

CH02-N10

Вариативность официального языка не может автоматически учитываться как ошибка представления.

CH02-N11

Идентичность материала, объём, время, компетенция или изменчивость рекомендаций не могут классифицироваться как различие в стиле без обоснования.

CH02-N12

Высокая стабильность сама по себе не является доказательством качества; следует также фиксировать постоянно неправильные результаты.

CH02-N13

Высокий средний показатель, если он сопровождается низким охватом, должен иметь общедоступные результаты с ограничением по охвату.

CH02-N14

Оценки подгрупп не следует публиковать как окончательные суждения о стране или языке при недостаточном объеме выборки и большой неопределенности.

CH02-N15

Уровни отказов и технических сбоев должны сохраняться как часть подлинного пользовательского опыта.

CH02-N16

Единственный счет, сделанный общедоступным, должен быть связан с каноническим распределением и записью области.

CH02-N17

Синтетические распределения не могут быть объединены с производительностью реального ИИ-продукта или реального объекта.

CH02-N18

Отчёт о распределении должен иметь ответственного человека или организацию.

26. ФОРМЫ ОШИБОК

CH02-F01— СРЕДНЕЕ МЫШЛЕНИЕ

Показан высокий глобальный средний показатель; язык, страна или критическая ошибка сохраняются.

CH02-F02— УДАЛИТЬ ПОДГРУППУ

Слабо работающие группы исключаются из отчета или объявляются «незначительными».

CH02-F03— КАРТА СФЕРЫ

Узкоспектральные хорошие результаты используются для всего мира, всех языков или всех продуктов.

CH02-F04— НОРМАЛИЗАЦИЯ СЕРЬЕЗНОСТИ

Незначительные недостатки и серьезные ошибки в области здоровья или лицензирования объединяются в один и тот же счет ошибок.

CH02-F05— СКРЫТЬ УХУДШЕНИЕ С ВРЕМЕННОЙ СРЕДНЕЙ

Падение или колебания становятся невидимыми только при усреднении за период.

CH02-F06— КУЛЬТ СТАБИЛЬНОСТИ

Предполагается, что последовательный ответ является автоматически правильным.

CH02-F07— ОШИБКА БУКВАЛЬНО ОДИНАКОВОГО ОТВЕТА

Для измерения согласованности используется буквальное совпадение текста вместо семантической корректности.

CH02-F08— УЧЁТ КОНТЕКСТУАЛЬНЫХ РАЗЛИЧИЙ КАК ОШИБКИ

Различия в соответствующих рекомендациях на основе реальных потребностей пользователя отмечаются как произвольная вариативность.

CH02-F09— УЧЁТ ПРОИЗВОЛЬНОЙ ВАРИАТИВНОСТИ КАК ПЕРСОНАЛИЗАЦИИ

Материальные противоречия при равных условиях пользователя оправдываются как «персонализация» без объяснения.

CH02-F10— ИЗМЕНЕНИЕ ЗНАМЕНАТЕЛЯ

Пострезультатный переход осуществляется между завершёнными, допустимыми или оценёнными наблюдаемыми знаменателями для повышения положительного исхода.

CH02-F11— ПОСТРЕЗУЛЬТАТНОЕ ВЗВЕШИВАНИЕ

Выбираются страна, язык или коэффициенты модели, обеспечивающие наивысший результат.

CH02-F12— ОБЪЕДИНЕНИЕ НЕСОВМЕСТИМЫХ ЯЧЕЕК

Различные условия запроса, продукта, времени или панели представлены как единое распределение.

CH02-F13— УДАЛЕНИЕ ОТКАЗОВ

Системные записи, на которые не было ответа, удаляются, чтобы получить высокую точность только среди тех, кто дал ответ.

CH02-F14— ОПРЕДЕЛЕННОЕ СУЖДЕНИЕ ИЗ МАЛОЙ ЯЧЕЙКИ

Оценка для страны или языка формируется на основе данных одного или нескольких пользователей.

CH02-F15— СКРЫТИЕ КРИТИЧЕСКИХ ОШИБОК

Редкие серьезные ошибки теряются в общей пропорции успешных результатов.

CH02-F16— ОТЧЕТ С ОДНОЙ ОЦЕНКОЙ

Публикуется только основная оценка без показа объема, неравенства, стабильности, неопределенности и серьезных ошибок.

CH02-F17— ВЫЧИСЛЕНИЕ РАСПРЕДЕЛЕНИЯ КАК «ВЕРОЯТНОСТИ ИИ»

Наблюдаемые результаты пользовательский интерфейс представляются так, как будто они являются неизменной и врожденной точкой зрения искусственного интеллекта.

CH02-F18— УЧЕТ НЕИЗМЕРЕННОГО КАК НУЛЕВОЙ ВАРИАЦИИ

Нетестированные области считаются гладкими и стабильными.

27. ПРОЦЕДУРА АУДИТА

Следует проверить, честно ли GEO-отчет отражает распределение, с помощью следующих шагов.

Шаг 1— Записать публичное заявление

Пример: «Глобальный NOMOS показатель бренда составляет 92». Полное предложение и интерфейс, на которой оно найдено, фиксируются.

Шаг 2— Определить целевое распределение

Для какого показателя:

- пользовательской выборки,
- ИИ-продуктов,
- стран,
- языков,
- периода времени

применяется?

Шаг 3— Извлечь измерительные ячейки

Определяется каждая комбинация материалов:

- Продукт
- Страна
- Язык
- Панель
- промпт
- Волна
- пользовательский интерфейс

Шаг 4— Воссоздать знаменатели

Назначенные задачи, выполненные задачи, действительные записи, рассмотренные записи и пройденные ответы должны учитываться как отдельные показатели.

Шаг 5— Извлечь статусы результатов

Не только уровень прохождения:

- частичный,
- неподдерживаемый,
- противоречивый,
- старый,
- Критический,

НЕИЗВЕСТНО,

отклоненный, неопределенный — распределение рассматривается.

Шаг 6— Проверить веса

Веса:

- установлены ли они заранее,
- на какой вопрос они отвечают,

изменились ли они после результата?

Шаг 7— Исследовать различия между подгруппами

Оцениваются лучшая группа, самая слабая достаточная группа, языковые различия, различия между странами, различия в продукте и различия панели.

Шаг 8— Проверка объема и недостатков

Какие области не были измерены? Какие ячейки недостаточны? Каковы причины недостатка? Могут ли отсутствующие данные быть систематическими?

Шаг 9— Проверить Определение

Оценивается изменение между Волны Сессий Пользователей Интерфейсов.

Шаг 10— Проверить Риск Очереди

Критические и другие серьезные ошибки:

- число,
- его частота,
- тип,
- эффект пользователя

оценивается отдельно в отношении

Шаг 11— Сравнить Публичное Резюме с Распределением

Отражает ли опубликованное для общественности отдельное число или предложение точно направление и риск полного распределения?

Шаг 12— Ограничить Результат

Публичное утверждение ограничено только уровнем, представленным распределением, и его охватом.

28. НЕОБХОДИМЫЕ ДОКАЗАТЕЛЬСТВА

В той степени, в которой это актуально для репрезентативного распределения или глобального показателя GEO, следует проверить следующие записи:

- Запись проверяемого объекта
- Определение целевой популяции
- Выборочная совокупность
- ИИ-продукт и пользовательские интерфейсы
- Ячейки страны и языка
- Запрос и версии
- Условия сессии и панели

- Волны измерений
- Назначенные номера задач
- Выполненные задачи
- Действительные и недействительные записи
- Исходные ответы
- Распределение статусов результатов
- Записи рассмотрения
- Записи критических ошибок
- Записи об отказах и неопределенных результатах
- Записи UNKNOWN
- Весы
- Заинтересованные стороны
- Результаты подгрупп
- Результаты стабильности
- Вне зоны охвата
- Расчеты доверия или неопределенности
- Публичная таблица показателей
- Ответственное лицо или организация
- Версия протокола и анализа
- История изменений

29. ЧЕК-ЛИСТ АУДИТА

Ясна ли целевая аудитория распределения? Указаны ли проверенные ИИ-продукты и пользовательские интерфейсы? Видны ли страна, язык, рамки запросов и времени? Раскрыты ли веса, использованные для расчета комплексного балла? Полностью ли определен знаменатель? Сообщаются ли недействительные записи отдельно? Сохраняются ли отказы и неопределенные результаты? Различаются ли UNKNOWN и NOT TESTED? Видны ли результаты подгрупп? Интерпретировались ли малые ячейки с чрезмерной уверенностью? Сохранена ли самая слабая материальная группа? Сообщаются ли критические ошибки отдельно? Показаны ли волны измерений отдельно? Скрывает ли среднее снижение или нестабильность? Различается ли лишь формальное различие от существенного различия по смыслу? Корректно ли классифицирована контекстно-зависимая вариация в рекомендациях? Объединены ли различные интерфейсы одного и того же провайдера без объяснения?

Включены ли в глобальный результат области, которые находятся вне рамок? Были ли разделены синтетические и реальные распределения? Отражает ли представленное публике предложение истинное направление полного распределения? Если один из существенных вопросов остался без ответа, результат распределения вызывает ограниченное доверие. Если без ответа остались более одного, один общий глобальный показатель не следует представлять как окончательное суждение публике.

30. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Общество хочет видеть один показатель».

Это верно. Один показатель может облегчить коммуникацию. NOMOS не полностью отвергает один показатель. Однако один показатель:

- объем охвата,
- Критическая ошибка,
- основа подгруппы,
- время,
- неопределенность

должен быть связан с информацией. Показатель может быть кратким. Он не может сократить правду.

Возражение 2— «Малые языки ненужно снижают глобальный показатель».

Глобальный вес населения может быть учтен для маленького языка с ограниченным весом. Однако сильное искажение представления в малом языке нельзя оставлять без внимания просто потому, что его доля в населении низка. Таким образом:

- основной показатель с учетом веса населения,
- в результате отдельной справедливости языка

Их можно публиковать вместе. Две метрики отвечают на разные вопросы.

Возражение 3— 'Если есть персонализация, разные ответы нормальны.'

Некоторые различия могут быть нормальными и правильными. Однако настройка:

- юридическое лицо,
- лицензия,
- ее текущая цена,
- реальная зона обслуживания

не должна изменяться произвольно. Персонализация может повлиять на надлежащее объяснение и советы для пользователя. Это не разрешение придумывать материальный факт.

Возражение 4— “Почему это было бы проблемой, если не дать один и тот же ответ?”

Проблема не в различии слов. Проблема в изменении материальной реальности. NOMOS не ищет точного текстового совпадения, а условную семантическую согласованность.

Возражение 5— «Почему имеет значение, если критическая ошибка возникает только один раз в ответах 1,000?»

Эта ошибка может быть редкой с точки зрения распространенности в популяции. Однако она может быть серьезной с точки зрения ущерба. Неконтролируемая потеря критической системы на самолете один раз на тысячу полетов не оценивается как всего лишь 0.1 процентов. В контексте GEO частоту и ущерб следует также рассматривать отдельно.

Возражение 6— «Мы не можем найти достаточную выборку для каждой маленькой страны или языка.»

Это правда. Поэтому:

- глобальная панель населения,
- панель наблюдателей по странам,
- панель языковой справедливости

можно установить отдельно. Наблюдение за одним человеком не превращается в оценку страны. Оно лишь указывает на наличие наблюдения. Правильный статус для недостаточной выборки:

НЕДОСТАТОЧНАЯ ВЫБОРКА

или:

НЕ ОЦЕНИВАЛОСЬ

должен быть.

Возражение 7— «Такое количество деталей слишком сложно для понимания клиентом».

Клиенту не обязательно видеть все формулы на главной странице. Однако ключевые факты, необходимые для принятия решения, нельзя скрывать. Краткая карточка с результатами должна отвечать на следующие вопросы: Каков общий результат? Каков самый серьезный риск? Какая область самая слабая? Что не было протестировано? Насколько надежен результат? На какую дату он относится? Сложность нужно управлять. Устранять её не следует.

Возражение 8— «Если одна и та же модель используется во всех странах, почему тогда будет различие между странами?»

Выход, видимый пользователю, может состоять не только из самого простого модели. Различия между странами включают:

- доступ к продукту,
- веб-ресурсы,
- локаль,
- языка,
- план,
- интерфейс,
- законодательство,
- персонализацию

Это может проявляться через него. Различие может не возникнуть. Однако отсутствие его нельзя предполагать без измерения.

Возражение 9— «Невозможно надежно оценивать, так как ответы ИИ изменчивы.»

Изменчивость не делает измерение невозможным. Напротив, это одна из характеристик, которую необходимо измерять. Оценка:

- конкретная дата,
- объем охвата,

- доверительный интервал
- информация о стабильности

может быть значимой, если она это несет. Однако утверждение о вечной уверенности не защищается.

Возражение 10— «Если промпт меняется, распределение полностью меняется.»

Это правильно. Следовательно, распределение зависит от промпта. GEO-1000 не даёт единственный ответ на все возможные вопросы. Предопределено:

- Основной зеркальный запрос,
- Регистр диагностических запросов

Измеряет специфические репрезентативные области. Структура промпта будет определена в Разделе 10.

ОБЩЕЕ ПРАВИЛО РАЗДЕЛА 32

Этот раздел не был написан против единственного балла. Он был написан с предположением, что один балл сам по себе представляет всю реальность. Людям нужны цифры. Учреждениям нужны рейтинги. Советы хотят видеть одну линию. Рынок может спросить: «Каков ваш NOMOS балл?» Этот вопрос законен. Но ответ не может быть только: «90.» Потому что за 90 могут существовать следующие реальности:

- Постоянная точность во всех языках
- 99 на некоторых языках, 45 на других
- Высокий средний показатель, но серьёзная критическая ошибка
- Высокая точность, но низкое покрытие
- Высокие мгновенные результаты, но быстрое падение
- Успех только в лёгких запросах
- Высокий показатель только среди оцениваемых ответов
- Положительное предположение о непроверенных странах
- Комбинирование выбранных продуктов ИИ

Значение оценки формируется видимостью основного распределения. Если распределение невидимо, оценка не вызывает доверия. Она лишь создаёт видимость уверенности. Следовательно, второй закон измерения NOMOS выглядит следующим образом:

> *Оценка GEO может быть сводкой распределения. / Она не может заменить распределение.*

Третий его закон таков:

> *Высокий средний показатель не снимает вопрос о том, кто слабо представлен.*

Четвёртый закон выглядит следующим образом:

> *Редкая серьёзная ошибка не является незначительной только потому, что она редка.*

Пятый закон звучит следующим образом:

> *Неисследованное поле не является положительной стороной распределения; это неизвестная граница.*

Порядок 2 из NOMOS

Не показывайте мне только среднее.

Покажите, кто ниже среднего.

Не скрывайте худший язык, показывая лучший язык.

Не используйте вес больших стран, чтобы стереть существование маленьких стран.

Не отправляйте одного человека в маленькую страну и затем объявляйте окончательный результат для этой страны.

Не преуменьшайте одну серьёзную ошибку только потому, что она скрыта в ответе, который на 99 процентов правильный.

Не объявляйте ошибкой только потому, что формулировка ответа изменилась. / Но когда изменяется сама реальность, не отвергайте это как простую стилистическую разницу.

Не выбрасывайте отклонения, неизвестные и непроверенные данные за пределы распределения.

Не меняйте вес, долю или охват после того, как увидели результат.

Сгенерируйте число. / Но не делайте страны, языки, пользователей, времена и ошибки, стоящие за числом, невидимыми.

Потому что представление сущности — это не среднее предложение. / Это распределение различных реальностей, встречающихся у разных людей.

Заключительное предложение главы

Объект измерения GEO — это не отдельный ответ или средний балл искусственного интеллекта; это совокупность представительных результатов, сгенерированных в определенной популяции пользователей при определенных условиях.

Нормативное ядро

Представление генеративной системы сущности ДОЛЖНО рассматриваться как условное распределение наблюдаемых результатов по определённым: - продуктам ИИ и пользовательским интерфейсам, - пользователям или единицам выборки, - странам, - языкам и локалям, - состояниям сессии и персонализации, - промптам, - и времени измерений. Один агрегированный показатель НЕ ДОЛЖЕН заменять раскрытие информации о: - охвате и покрытии, - распределении результатов, - вариации по подгруппам, - частоте критических ошибок, - неизвестной и непроверенной массе, - временной стабильности, - правилах взвешивания, - и неопределенности. Равные средние значения НЕ ДОЛЖНЫ интерпретироваться как одинаковое качество представления, если их распределения по подгруппам, степени тяжести, охвату или стабильности существенно различаются. Неисследованные группы НЕ ДОЛЖНЫ учитываться как соответствующие, а редкие высокоопасные исходы ДОЛЖНЫ оставаться отдельно видимыми от средней производительности.

ГЛАВА 03

Что именно является объектом аудита?

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K10; K11

Граница главы

Раздел 1 установил следующее положение:

> *Одиночный ответ не является GEO баллом.*

Раздел 2 добавил следующее положение:

> *Представление сущности в генеративных системах — это распределение, которое встречается среди пользователей, продуктов ИИ, стран, языков, сеансов и времени.*

Этот раздел отвечает на следующий обязательный вопрос:

> *Чье или какое представление точно измеряет это распределение?*

Компания? Бренд? Доменное имя? Продукт? Услуга? Основатель? Группа компаний? Местная юридическая сущность в определённой стране? Франшизный бизнес? Когда система искусственного интеллекта видит только домен apple.com:

- доменное имя,
- бренд,
- материнская компания,
- местная торговая организация,
- семейство продуктов,
- сервис магазина

Будет ли система рассматривать их всех как одну и ту же сущность? Если различие не будет уточнено до начала аудита, результат может быть рассчитан правильно, но применен к неверному объекту. Эта глава:

- аудированная сущность,
- якорь идентичности, используемый в запросе,
- сущность, проанализированная системой,
- единица, к которой будет применено решение о соответствии,
- взаимоотношения между сущностями,
- какой сущности принадлежат атрибуты,
- временные и географические границы идентичности

определяет. В этой главе пока не рассматривается:

- как будут проверяться все утверждения о сущности,
- как арбитры будут оценивать атомарные утверждения,
- квоты по странам и языкам,
- окончательная формула оценки NOMOS

не завершена. Задача Раздела 3 более фундаментальна:

> *Зафиксировать идентичность и границы объекта, который измеряется, до начала измерения.*

NOMOS Вызов

Вы задаёте мне этот вопрос: «Что это за компания Apple.com?» В вопросе может быть как минимум три разных объекта:

- доменное имя apple.com
- бренд Apple
- юридическая компания, связанная с доменным именем и брендом

Я бы выбрал один из этих вариантов. Возможно, правильный. Возможно, неправильный. Возможно, я объединю несколько в одном ответе. Затем вы оцените ответ. Но вы заранее не сказали, какую сущность вы оцениваете. Я могу:

- описать домен как интернет-магазин,
- описать бренд как бренд потребительской электроники,
- описать юридическую компанию как глобальную корпорацию,

Я могу перечислить продукты так, как если бы компания сама их представляла. Мой ответ может выглядеть связным. Но если объект аудита не ясен, то вопрос, правильный он или нет, также остается неопределенным. Теперь рассмотрим более сложный пример.

Брендом управляет материнская компания в одной стране; дочерняя компания — в другой; а независимый дистрибьютор — в третьей стране. Пользователь спрашивает: «Предоставляет ли этот бренд услуги в Турции?» Какая организация должна отвечать? Глобальный сайт бренда? Местная компания в Турции? Дистрибьютор?

Оператор франшизы? Лицензия, принадлежащая глобальной компании, автоматически передается местной франшизе? Отзывы клиентов о местной компании доказывают ли глобальное качество основного бренда? Можно ли использовать количество сотрудников материнской компании как показатель возможностей местной компании?

Можно ли перенести награду продукта на все компании, которые его продают? Становится ли личный опыт основателя корпоративным возрастом компании? Считается ли предыдущий и новый владелец доменного имени одним и тем же юридическим лицом?

Когда компания приобретает другую компанию, становятся ли все достижения предыдущей компании характеристикой новой компании? Балл GEO может быть сгенерирован без ответа на эти вопросы. Однако неизвестно, что представляет собой этот балл. Первое постановление этого раздела следующее:

> *Сущность не может быть измерена, пока она не будет определена.*

Второе положение гласит:

> *Наличие одинакового названия не означает, что это одна и та же сущность.*

Её третье положение гласит:

> *Отношение не позволяет автоматически переносить свойства между сторонами.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела заключается в определении идентичности, границ и отношений измеряемой сущности в GEO-1000 до того, как будут видны результаты. Раздел делает следующие нормативные различия:

- Доменное имя и юридическая компания
- Бренд и владелец бренда
- Материнская компания и дочерняя компания
- Группа и компания группы
- Продукт и производитель
- Сервис и поставщик услуг
- Основатель и учреждение
- Сотрудник и учреждение
- Держатель лицензии и сторона, использующая бренд
- Франчайзер и оператор франшизы
- Производитель и дистрибьютор
- Платформа и продавец на платформе
- Поставщик технологий и пользователь технологий
- Клиент и деловой партнер
- Активные отношения и исторические отношения
- Официальное название и описательный перевод
- Разные объекты с одинаковым названием
- Предыдущие и последующие версии объекта
- Аудируемый объект и объект, к которому будет применено решение о соответствии

В конце этого раздела каждый аудит GEO должен быть способен предоставить единый и отслеживаемый ответ на следующий вопрос:

> *Какому именно субъекту, с какими отношениями, за какой период и для каких цифровые каналы принадлежит этот результат и решение?*

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

> *Перед началом сбора данных каждый аудит GEO-1000 должен определить уникальный объект аудита, каноническую запись объекта, границы области, временные границы и граф отношений.*

Ни один из следующих элементов сам по себе недостаточен для идентификации объекта:

- название,
- логотип,
- доменное имя,
- учетная запись в социальной сети

Эти идентификаторы сами по себе не определяют объект аудита. Аудируемый объект должен быть описан, при необходимости, как минимум через следующие элементы:

- Тип субъекта
- Каноническое название

- Юридическое название
- Торговые марки
- Официальные локализованные названия
- Аббревиатуры и псевдонимы
- Дата основания или начала деятельности
- Юрисдикция
- Регистрационный номер или эквивалент официального удостоверения личности
- Отношения материнской и дочерней компании
- Право собственности на товарные знаки
- Контроль за доменным именем
- Право собственности на продукты и услуги
- Ответственный за фактическое оказание услуг
- Истец
- Обладатель доказательств
- Активные страны и регионы
- Срок действия
- Исторические названия и отношения
- Интерфейсы, включенные в аудит
- Сущности и интерфейсы, не входящие в область охвата

Основной GEO балл не должен публиковаться, пока эта запись не будет завершена.

3. ЧТО ТАКОЕ СУБЪЕКТ?

В этом протоколе субъект — это объект, которому можно присвоить отдельную и отслеживаемую идентичность, характеристику, отношение, ответственность или временной статус. Субъект может быть:

- человеком,
- юридическим лицом,
- государственным учреждением,
- брендом,
- продукта,
- услуги,
- доменное имя,
- цифровым приложением,
- группой компаний,
- физическим подразделением,
- программой,
- свидетельством,
- публикацией,
- исследовательским проектом

Это возможно. Не каждый из перечисленных элементов должен быть отдельным субъектом. Однако, если одна из следующих областей существенно изменяется, может потребоваться отдельная запись о субъекте:

- Юридическая ответственность
- Собственность
- Лицензия
- Поставка товаров или услуг
- Юрисдикция
- Период времени
- Обязательство перед пользователем
- Доказательства и право собственности на утверждения
- Сфера аудита

4. Шесть отдельных объектов в GEO-аудите

В GEO-аудите слово «сущность» часто скрывает шесть отдельных функций. Эти шесть объектов необходимо различать друг от друга.

4.1. Якорь запроса

Якорь запроса — это выражение или знак, используемый для вызова существования в запросе пользователя. Примеры:

- Название компании
- Название бренда
- Доменное имя
- Имя основателя
- Название продукта
- Название приложения
- Название местной компании
- Аббревиатура

Пример: «Что это за компания Apple.com?» якорь запроса здесь: apple.com является доменным именем. Якорь запроса не обязательно должен быть самим субъектом. Это только точка входа, используемая системой для определения субъекта.

4.2. Целевой субъект

Целевой субъект — это субъект, который протокол на самом деле хочет измерить. Пример: канонический корпоративный субъект, связанный с доменным именем apple.com, может быть выбран в качестве цели. Однако его точная идентичность должна быть зафиксирована до аудита. Целевой субъект не может быть выбран на основании результатов после запроса.

4.3. Субъект, определенный системой

Определенный субъект — это то, как продукт ИИ интерпретирует якорь запроса как субъект. Представление:

```
E_i
```

может быть использовано. Целевой субъект:

E*

является оценкой разрешения идентичности, которая задает фундаментальный вопрос:

```
E_i= E*?
```

Однако равенство не всегда означает просто равенство по названию. Система может определить правильный бренд и корпоративное образование без указания правильного юридического лица. Поэтому последующее сопоставление идентичности:

- точное совпадение,
- приемлемое совпадение бренда,
- частичное совпадение,
- неверная сущность,
- неопределенно,
- слияние нескольких лиц

может быть классифицировано как.

4.4. Объект иска

Объект иска — это лицо, к которому принадлежит конкретная характеристика в ответе. В рамках одного ответа:

- дата основания головной компании,
- характеристика продукта к продукту,
- отзыв клиента для местного поставщика услуг,
- лицензия на конкретное юридическое лицо,
- опыт основателя к человеку

может принадлежать. Даже если ответ начинается с правильного целевого объекта, он может переносить признаки на неправильные объекты. Поэтому каждое фактическое утверждение следует оценивать по следующей структуре:

```
(субъект, отношение или характеристика, значение, время, объем)
```

Пример: «X имеет лицензию Y». Здесь следует определить, к чему это относится:

- юридическим лицом,
- страны,
- в какие даты,
- для деятельности

это принадлежит.

4.5. Объект аудита

Объект аудита — это определённая сущность или пакет сущностей, в рамках которых одновременно исследуются доказательства, содержание, техническая интерфейс и результаты работы ИИ. Объектом аудита может быть отдельная сущность. Пример: конкретная юридическая компания. Либо это может быть пакет с чёткими границами: основной бренд + каноническое доменное имя + два указанных продукта + турецкая и английская публичные интерфейсы. Если используется пакет, сущности внутри него должны сохранять свои отдельные идентичности. "Пакет" не означает, что все атрибуты передаются друг другу.

4.6. Единица соответствия

Единица соответствия — это конкретный объект, к которому юридически и нормативно будет применено решение аудита. Решение о соответствии:

- применяется ко всей группе компаний,
- применяется ко всем языкам,
- применяется ко всем доменным именам,

- ко всем продуктам

не может применяться автоматически. Пример: «Публичные страницы бренда NobleJackal на сайте example.com на турецком и английском языках признаны соответствующими указанному стандарту и в указанный период». Это решение:

- все внутренние процессы компании,
- аффилированные компании,
- отдельные приложения,
- языки, которые будут добавлены в будущем,
- продукты, находящиеся вне рамок

не охватывает автоматически.

5. ТИПЫ СУБЪЕКТОВ

Каждое аудируемое предприятие должно быть классифицировано с одним или несколькими типами. Можно различать первичный и вторичный типы.

5.1. Юридическое лицо

Это компания, которая обладает отдельными правами, обязанностями или регистрацией согласно закону или соответствующему правовому регламенту. Ее поля в той мере, в которой они релевантны:

- Юридическое название
- Тип компании
- Место регистрации
- Регистрационный номер
- Налоговый или эквивалентный идентификатор
- Дата учреждения
- Активный статус
- Уполномоченный представитель
- Отношения материнской и дочерней компаний

Юридическое лицо может носить то же название, что и бренд. Однако это не один и тот же объект.

5.2. Бренд

Это имя, используемое на рынке для продукта, услуги или корпоративной идентичности. Бренд:

- может не иметь отдельного юридического лица,
- может принадлежать другой компании,
- может использоваться по лицензии,

может управляться несколькими компаниями в разных странах. Следует также указать фактическую организацию, ответственную за работу, заявленную брендом.

5.3. Группа компаний

Это структура, сформированная несколькими юридическими лицами, которые имеют совместное владение или контрольные отношения. «Группа» не может автоматически передавать всем членам следующие характеристики:

- Лицензия
- Сертификат
- Количество клиентов
- Количество сотрудников
- Выручка
- Офис
- Результат аудита
- Статус безопасности
- Знак соответствия

Если значение на уровне группы действительно консолидировано, это должно быть ясно указано.

5.4. Дочерняя компания или Аффилированное лицо

Юридическое лицо, находящееся во владении или под контролем другого юридического лица. Дочерняя компания:

- может быть отдельной стороной по контракту,
- отдельным владельцем лицензии,
- отдельным работодателем,
- отдельным контролером данных

может быть. Характеристики материнской компании по умолчанию не передаются дочерней компании.

5.5. Доменное имя

Это цифровой адрес и интерфейс публикации. Доменное имя:

- не является юридическим лицом,
- не предоставляет услуги самостоятельно,
- не обязано владеть товарным знаком,

может контролироваться разными владельцами или операторами в определенные моменты времени. Доменное имя может быть отправной точкой аудита. Однако также следует указать, под каким юридическим лицом было принято решение.

5.6. Веб-сайт или цифровое издание

Это совокупность конкретного контента и приложений на доменном имени. Одно и то же доменное имя может быть:

- корпоративный сайт,
- магазин,
- портал поддержки,
- страница для инвесторов,
- документация для разработчиков,
- разные версии для стран и языков

могут содержать. Владельцы контента и объемы контроля за ними могут различаться.

5.7. Продукт

Коммерческий или технический объект, выполняющий конкретную функцию, модель или версию. У продукта:

- цена,
- безопасность,
- техническая характеристика,
- свидетельством,
- подходящесть для пользователя

не могут использоваться как общие характеристики, принадлежащие компании. Версии внутри одной линейки продуктов также могут отличаться.

5.8. Сервис

Деятельность, предлагаемая в рамках конкретного объема, для пользователя, страны, метода доставки и ответственности. Услуга:

- может быть реализована брендом,
- может быть оплачена другим юридическим лицом,

может быть предоставлена независимым партнером. Аудит должен различать, кто:

- продал её,
- доставил её,
- гарантировал её,
- юридически принял на себя

должно быть отделено.

5.9. Человек

является основателем, руководителем, сотрудником, консультантом, экспертом или публично признанным представителем. Образование, опыт и мнение человека:

- образование,
- опыт,
- лицензия,
- мнение

не автоматически переходят к учреждению. Точно так же награда или сертификат учреждения не считается профессиональной компетенцией человека.

5.10. Филиал или физическое местоположение

Это физическая точка обслуживания, которая принадлежит основному учреждению или управляется отдельно. Филиал:

- часть материнской компании,
- отдельное юридическое лицо,
- франшиза,
- независимый лицензионный оператор

возможно. Отношение по позиции не показывает исключительно юридическую связь.

5.11. Франшиза

Это может быть независимый бизнес, работающий на основании права использования бренда или лицензии на бизнес-модель. Франчайзер должен:

- репутацию,
- контроль,
- обучение

не гарантирует автоматически каждое действие франчайзи. Локальная ошибка франчайзи не может автоматически распространяться на весь глобальный бренд. Отношение должно быть указано с его охватом.

5.12. Дистрибьютор, дилер или продавец

Сторона, которая продает или распространяет продукт или услугу на конкретном рынке, независимо или в связке с другими. Лицензия производителя, контроль или корпоративное соответствие автоматически не передаются дистрибьютору. Производительность службы поддержки продавца также не является тем же показателем, что и качество продукции производителя.

5.13. Платформа и продавец на платформе

Рынок или платформа и независимые продавцы на платформе являются отдельными субъектами. Платформа:

- платеж,
- листинг,
- техническая инфраструктура

могут быть предоставлены. Продавец:

- продукта,
- доставка,
- гарантия,
- взаимоотношения с клиентами

может нести ответственность. Если ИИ объединяет эти роли, может произойти неправильное распределение ответственности.

5.14. Программа, Значок или Сертификат

Предлагается учреждением:

- учебная программа,
- знака соответствия,
- свидетельством,
- членство,
- значок

Он может содержать отдельные записи об объектах. Учреждение, которое публикует программу, не обязательно является тем же учреждением, которое участвует в программе. Особенность значка не переносится на все характеристики учреждения, которое его носит.

6. ГРАФ ОБЪЕКТОВ

Одной записи об объекте может быть недостаточно для объяснения сложных отношений. По этой причине для каждого аудита должен быть создан граф объектов. Представления:

G_E = (V, R)

Здесь:

- V: узлы объектов
- R: направленные отношения между узлами

Отношения должны определяться явными глаголами.

6.1. Основные типы отношений

В той мере, в какой это уместно, могут использоваться следующие отношения:

- owns — владеет
- ownedBy— принадлежит
- controls — контролирует
- controlledBy— контролируется
- operates — управляет
- operatedBy— управляется
- publishes — публикует
- publishedBy— публикуется
- производит — производит
- manufacturedBy— производится
- providesService— предоставляет услуги
- serviceDeliveredBy— услуги предоставляются
- licencesBrandTo— лицензирует бренд для
- franchisesTo— продается по франшизе
- distributesFor— распространяется для
- sells — продаёт
- employs — нанимает
- foundedBy— основана
- memberOf— является членом
- certifiedBy— сертифицировано по указанной области
- auditedBy— проверено аудитом
- partneredWith— имеет партнерство в указанной области с
- customerOf— является клиентом
- supplierTo— является поставщиком для
- formerlyKnownAs— ранее известен как
- predecessorOf— является предшественником
- successorTo— является преемником
- acquiredBy— был приобретен
- mergedInto— был объединен с
- ceasedOn— прекратил деятельность

Неопределенные глаголы вроде «связан с», «партнер» или «работают вместе» следует по возможности не использовать самостоятельно.

6.2. Направление отношения

Отношения являются направленными. «A предоставляет технологию B» — это не то же самое, что: «B предоставляет технологию A». Бренд может принадлежать компании. Компания не принадлежит бренду. Клиент мог работать с агентством. Агентство не принадлежит клиенту. Если направление не указано, система может неправильно передать ответственность и полномочия.

6.3. Время отношений

Каждое финансовое отношение в той мере, в которой это имеет значение:

- дата начала,
- дата окончания,
- статус активности,
- дата последней проверки

должны быть указаны. Старая партнерская связь не является активной. Бывший сотрудник не является текущим руководителем. Дистрибьютор, когда-то использовавшийся, может не являться уполномоченным представителем сегодня.

6.4. Объем отношений

Партнерство:

- может быть ограничено определенным продуктом,
- страны,
- кампания,
- дата,
- событие

Нельзя обобщать, что сотрудничество в рамках одного проекта применяется ко всем корпоративным действиям.

7. ПРЕДПОЛАГАЕМОЕ ПЕРЕДАНИЕ ХАРАКТЕРИСТИК ЗАПРЕЩЕНО

Наличие отношений между организациями не автоматически передаёт характеристику от одной к другой. Один из основных принципов этого раздела:

> *Передача характеристик запрещена по умолчанию; однако она может быть разрешена при явной связи, охвате, сроках и наличии доказательств.*

7.1. Передача лицензии

Лицензия материнской компании:

- дочерней компании,
- франшизе,
- дистрибьютору,
- бренду

не передавать автоматически. Необходимо проверить официальный объем лицензии.

7.2. Передача сертификата

Сертификат продукта не может быть передан всей продуктовой линейке. Сертификат качества одного офиса не может быть распространен на глобальную группу. Решение компании о соответствии не подтверждает личную квалификацию её сотрудников.

7.3. Передача числа сотрудников

Общее число сотрудников группы не может использоваться как число сотрудников местной компании. Сеть агентств или пул фрилансеров нельзя показывать как число прямых сотрудников.

7.4. Передача числа клиентов

Число клиентов основной группы не может быть указано как число клиентов недавно созданного бренда. Количество пользователей на платформе может не соответствовать числу активных клиентов конкретного продукта.

7.5. Передача даты основания

Год начала работы основателя в отрасли не является датой основания компании. Компания не может использовать дату основания старого бренда, который она приобретает, в качестве собственной юридической даты основания. Дата бренда и дата юридического лица должны быть указаны отдельно.

7.6. Успешное завершение и передача случая

Проект, выполненный сотрудником на предыдущем месте работы, не является корпоративным кейсом новой компании. Успех приобретенной компании без объяснения интеграции и масштаба не считается прямым успехом новой группы.

7.7. Передача права на участие

Право на участие, предоставленное для определенных языковых страниц доменного имени:

- для всей группы компаний,
- для новых продуктов,
- для других доменных имен,
- для будущих страниц

не передается.

8. ЧЕТЫРЕ ОТДЕЛЬНЫЕ СЛОЯ СОБСТВЕННОСТИ

В цифровом представлении слово «владелец» может иметь несколько значений. Следует различать четыре уровня.

8.1. Законное владение

Сущность является владельцем по закону или официальной регистрации.

8.2. Владение товарным знаком

Это владелец прав на товарный знак или торговую идентичность.

8.3. Цифровое управление интерфейсом

Это сторона, которая фактически контролирует доменное имя, веб-сайт, социальный профиль или приложение. Юридический владелец и технический контролер могут быть разными.

8.4. Содержание и право собственности на утверждение

Это сторона, ответственная за точность опубликованного заявления о материале. Агентство может управлять сайтом. Однако истинным владельцем утверждения на цену или лицензию может быть клиентское учреждение. Технический поставщик не становится владельцем материала утверждения только потому, что он опубликовал содержание.

9. КТО ПРОДАЁТ УСЛУГУ, КТО ЕЁ ОКАЗЫВАЕТ?

Это одно из самых важных различий сущностей в оценке GEO. В цепочке предоставления услуг следующие стороны могут различаться:

- Тот, кто использует бренд
- Тот, кто рекламирует
- Тот, кто готовит предложение
- Тот, кто подписывает контракт
- Тот, кто выставляет счёт
- Тот, кто фактически предоставляет услугу
- Тот, кто предоставляет гарантию
- Тот, кто обрабатывает данные
- Тот, кто рассматривает жалобу
- Тот, кто несёт юридическую ответственность

ИИ может назначить все эти роли одной стороне, лишь указав название бренда. Эта ошибка особенно распространена в:

- туризме,
- финансах,
- здравоохранении,
- консалтинг,
- франшиза,
- платформа,
- агентство,
- образование

сферы деятельности могут привести к финансовым последствиям. Каждая запись об услуге должна отвечать на следующие вопросы: С кем пользователь заключает контракт? Кто получает оплату? Кто выполняет услугу? Кто несет профессиональную или юридическую ответственность? Кто отвечает за гарантийные и возвратные обязательства? Какова роль бренда? Есть ли поставщики третьей стороны? Это различие также определяет, к какой организации применяется оценка GEO.

10. ГРАНИЦА СУЩНОСТИ

Граница сущности — это линия, которая определяет, какие функции, отношения и интерфейсы включает в себя проверяемый объект и какие нет. Граница сущности может быть определена на четырех уровнях.

10.1. Основная идентичность

Это фундаментальная идентичность сущности, которая должна оставаться неизменной:

- Каноническое название
- Тип субъекта
- Юридическая или институциональная идентичность
- Основная деятельность
- Первичное владение
- Активный статус

10.2. Контролируемая среда

Непосредственно контролируется организацией:

- доменные имена,
- приложения,
- бренды,
- продукты,
- документы,
- официальные профили

Это поле. Контроль не означает, что все функции сливаются в единую идентичность.

10.3. Связанная среда

Филиалы — местные компании, франшизы и контролируемые партнёры. Они сохраняют свои отдельные идентичности.

10.4. Среда внешних отношений

Клиенты, поставщики, СМИ, членства, поставщики технологий и независимые партнёры. Сущности в этой среде не являются частью проверяемой организации. Они являются объектами отношений.

11. ЗАКРЫТИЕ ГРАНИЦЫ ОРГАНИЗАЦИИ БЕЗ ПРОСМОТРА РЕЗУЛЬТАТОВ

Объект аудита не может быть расширен или сужен на основе результатов. Следующее поведение запрещено:

- Добавление положительных результатов группы к бренду
- Исключение неудачной аффилированной компании из сферы контроля
- Распространение положительного результата продукта на всю продуктовую линейку
- Объявление плохой языковой версии отдельной сущностью

- Оставление критической ошибки локальной операции при создании видимости чистоты глобального бренда
- Предъявление сильных локальных доказательств как доказательств глобальной компании

Сфера действия сущности:

- Перед началом аудита,
- Перед распространением промптов,
- До того, как будут видны результаты

Она должна быть сначала версионирована.

12. КАНОНИЧЕСКАЯ ЗАПИСЬ СУЩНОСТИ

Для каждой проверяемой организации должен быть создан Канонический Запись Сущности. Эта запись уточняет, какой объект оценивают арбитры и системы.

12.1. Минимальные поля

Идентификация

NOMOS ID сущности Каноническое имя Тип сущности Юридическое название Аббревиатуры Официальные локальные названия Бывшие названия

Юридический и корпоративный статус

Юрисдикция Регистрационный номер Дата основания Активный статус Материнская компания Дочерние компании

Бренды и продукты

Собственные бренды Лицензированные бренды Продукты Услуги Версии или модели

Цифровые каналы

Канонические доменные имена Локальные доменные имена Приложения Официальные профили Машиночитаемые записи

Ответственность

Истец Владелец контента Владелец технической интерфейсы Владелец предоставления услуг Юридически ответственный

Время

Версия записи Дата начала действия Дата окончания действия Дата последней проверки Исторические изменения

Область

Аудированные страны Аудированные языки Аудированные продукты Сущности и интерфейсы вне области аудита

12.2. Статусы сущностей

Каждая запись сущности должна иметь один из следующих статусов:

АКТИВНО

НЕАКТИВНО

ИСТОРИЧЕСКОЕ

ЗАМЕНЕНО

ОБЪЕДИНЕНО

ПРИОБРЕТЕНО

РАСПУЩЕНО

ПРОВИЗОРНОЕ

СПОРНОЕ

НЕИЗВЕСТНО

НЕИЗВЕСТНАЯ или СПОРНАЯ личность не может быть окончательно разрешена незаметно.

13. УРОВНИ ПОДТВЕРЖДЕНИЯ ИДЕНТИЧНОСТИ

Подтверждение идентичности записи сущности можно классифицировать в зависимости от используемых источников.

ID-0— Личность неизвестна

Существует только имя или доменное имя. Юридическое или организационное соответствие не проверяется.

ID-1— Утверждение первой стороны

Учреждение идентифицирует себя определённым образом. Это ценно. Это не независимая проверка.

ID-2— Поддерживаемая организационная идентичность

Записи первой стороны, доменное имя и другие надёжные записи согласованы друг с другом.

ID-3— Проверенная личность с официальной записью

Юридическое лицо было проверено с использованием соответствующих официальных или надежных записей.

ID-4— Личность, проверенная по различным источникам и связям

Идентичность, бренд, продукт, доменное имя и корпоративные отношения согласованы в нескольких надёжных записях.

ID-5— Идентичность, поддерживаемая со временем

Личность и связи проверялись в разные периоды, а история изменений сохраняется. Эти уровни не являются окончательной оценкой соответствия. Они указывают на степень доверия к записи о сущности.

14. РАЗНЫЕ СУЩНОСТИ С ОДИНАКОВЫМ ИМЕНЕМ

Могут существовать разные сущности с одинаковыми или похожими именами. Эта ситуация создаёт столкновения сущностей. Примеры различий:

- Компании с одинаковым названием
- Учреждения, использующие одинаковую аббревиатуру
- Один и тот же бренд в разных странах
- Старая и новая компания
- Название бренда с личным именем
- Название компании с названием продукта
- Бренд с общим словом

Аудит не может основываться только на совпадении имен.

14.1. Поля разрешения коллизий

Юрисдикция Тип юридической компании Доменное имя Физический адрес Дата основания Основатель или руководитель Продукт и услуга Владельец бренда Регистрационный номер Название на местном языке

14.2. Двусмысленное утверждение

Если утверждение относится к нескольким разумным субъектам, есть три варианта: Утверждение уточняется. Классифицируется отдельно как тест разрешения субъекта. Ожидается, что система запросит уточнение. Полученный ответ на двусмысленное утверждение нельзя тихо добавлять к основному показателю точности.

15. СУБЪЕКТ, НЕСКОЛЬКО ДОМЕННЫХ ИМЕН

Компания:

- глобальный домен,
- локальные домены страны,
- домены продуктов,
- домен инвестора,
- домен поддержки,
- старые домены

может использовать. Не все эти домены автоматически подпадают под одну и ту же область аудита. Для каждого домена:

- контроллер,
- владелец контента,
- целевая аудитория,
- языка,

- объем продукта,
- статус активности

Это необходимо определить. Правильная запись в одной области не исправляет автоматически ошибку в другой области.

16. ДОМЕННОЕ ИМЯ, БОЛЕЕ ОДНОЙ СУБЪЕКТА

На одном доменном имени:

- материнская компания,
- различные бренды,
- местные компании,
- продукты,
- партнеры,
- продавцы

Это может иметь место. В этом случае собственность субъекта должна определяться на основе URL или семейства контента. Пример:

Интерфейс	Роль существования
Главная страница	Глобальный бренд
Инвесторские связи	Юридическая материнская компания
Страница с локальными ценами	Локальная организация продаж
Маркетплейс	Независимые продавцы
Страница карьеры	Указанные работодатели
Портал поддержки	Оператор продукта или услуги

Наличие одного доменного имени не означает единственной ответственности.

17. ВРЕМЕННАЯ ИДЕНТИЧНОСТЬ СУБЪЕКТА

Идентичность субъекта может со временем изменяться. Примеры случаев:

- Изменение названия компании
- Бренд передан другой компании
- Слияние
- Поглощение
- Деление
- Прекращение деятельности
- Изменение доменного имени
- Передача продукта другой компании
- Уход основателя
- Смена местного оператора

По этой причине регистрация сущности зависит от времени. Её можно рассматривать как E(t).

17.1. Историческая идентичность и текущая идентичность

Отношения, которые были корректны в прошлом, не могут переноситься в настоящее. Пример: «X был управляем Y в 2022.» — это историческая информация. Следующее предложение отличается: «X управляется Y.»

17.2. Передача функций после приобретения

Приобретение:

- собственность,
- некоторые субъекты,
- могут передавать определённые контракты

. Тем не менее:

- все исторические достижения,
- все сертификаты,
- все отношения с клиентами,
- все лицензии

не передаются автоматически новому субъекту. Необходимо разъяснить, какие функции передаются юридически и фактически.

17.3. История доменного имени

Доменное имя может нести следы контента или ссылки от предыдущего владельца. Новый владелец не считается автоматически получившим прежние полномочия. Если система ИИ объединяет старые и новые сущности, возникает путаница в исторических сущностях.

18. МНОГОЯЗЫЧНАЯ ИДЕНТИЧНОСТЬ СУЩНОСТИ

Имена сущности на разных языках должны быть связаны с одной и той же личностью. Однако не все переводы являются официальными именами. Необходимо различать следующие типы:

- Юридическое название
- Название бренда
- Официальное локализованное имя
- Транслитерация
- Описательный перевод
- Аббревиатура
- Бывшее имя
- Общеупотребительное имя

Пример: описание бренда на турецком языке не обязательно должно быть юридическим или официальным турецким именем бренда. Машинная запись должна четко указывать эти типы.

19. СТРАНА И ЮРИСДИКЦИЯ МОГУТ ИЗМЕНИТЬ ГРАНИЦУ СУЩНОСТИ

Под глобальным брендом могут существовать разные юридические или операционные структуры в разных странах. Когда пользователь спрашивает: «Осуществляет ли компания X деятельность в Германии?», ответ должен различать:

- наличие глобального бренда,
- местную юридическую структуру,
- отношения с дистрибьютором,
- возможность предоставления удалённых услуг

«Бренд известен в стране» не то же самое, что «Местная юридическая компания предоставляет услуги».

19.1. Роли организаций в разрезе стран

Для каждой страны, применительно к ситуации:

- Юридический продавец
- Выставитель счетов
- Контролёр данных
- Поставщик услуг
- Сопровождающий / Поддерживающий
- Дистрибьютор
- Франчайзи
- Держатель лицензии
- Представитель бренда

должно быть определено.

19.2. Общая языковая страница не является доказательством местной компании

Наличие веб-сайта на немецком языке не доказывает, что существует юридическое лицо или физический офис в Германии. Английская страница не создает глобальную юридическую юрисдикцию. Язык — это не юрисдикция.

20. СИНТЕТИЧЕСКОЕ APPLE.COM УКАЗАНИЕ ПРИСУТСТВИЯ

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Следующий пример подготовлен исключительно для демонстрации метода разрешения сущностей. Это не является фактичесальным аудитом Apple Inc., текущим обзором корпоративной структуры или реальным результатом какого-либо продукта ИИ. Запрос: «Что за компания Apple.com?»

20.1. Якорь промпта

apple.com

20.2. Кандидатные цели

Доменное имя и веб-сайт Бренд Apple Корпоративное образование Canonical Функция интернет-магазина Семейство продуктов Если цель не определена до начала аудита, судьи могут оценивать разные объекты.

20.3. Предлагаемый фиксация целевого объекта

Цель для синтетического тестирования может быть определена следующим образом: «основной корпоративный технологический объект, связанный с каноническим доменным

именем apple.com, представляющий идентичность компании и ее основные виды деятельности.» Это определение:

- не включает автоматически технические характеристики отдельных продуктов,
- местные филиалы компании в каждой стране,
- цены в магазинах,
- инвестиционные рекомендации

автоматически.

20.4. Синтетический ответ A

«Apple.com— это глобальный корпоративный веб-сайт, связанный с технологическими продуктами и цифровыми услугами.» Оценка: Это не полностью соответствует учреждению с данным доменным именем. Идентичность компании может быть представлена не полностью. Основная деятельность в целом может быть правильной. Может отсутствовать материальная идентичность.

20.5. Синтетический ответ B

«Apple.com— это официальный веб-сайт компании под названием iPhone.» Оценка: продукт и компания объединены. Целевая сущность определена неправильно. Имеется ошибка идентичности.

20.6. Синтетический ответ C

“Apple.com— это только независимый магазин, который продает телефоны онлайн.” Оценка: Функция доменного имени чрезмерно сужена. Корпоративная сущность и сфера деятельности искажены. Характеристика независимого магазина может быть неверной.

20.7. Синтетический ответ D

«Apple.com является одной из основных корпоративных и продуктовых информационных платформ бренда Apple; юридическое лицо, управляющее сайтом, и местная торговая сторона также должны определяться в зависимости от страны.» Оценка: доменное имя, бренд и местная юридическая роль разделены. Анализ целевого субъекта более дисциплинирован. Другие существенные утверждения ответа должны быть подтверждены отдельно. Этот пример показывает:

> *Ответ может сосредоточиться на неверная сущность, даже если он содержит много правильных слов.*

Точность идентичности важнее точности содержания.

21. СЛОЖНЫЙ СИНТЕТИЧЕСКИЙ СЛУЧАЙ

СИНТЕТИЧЕСКИЙ СЛУЧАЙ — НЕ РЕАЛЬНОЕ УЧРЕЖДЕНИЕ

Рассмотрим глобальный бренд под названием Asteron. Структура:

- Asteron Holdings Ltd. — материнская компания
- Asteron Travel — бренд
- Asteron Turkey Tourism Ltd. — местная компания
- Mira Destination Services — независимый местный партнер по операциям
- Asteron Club Antalya — франчайзинговое предприятие
- asteron.example — глобальное доменное имя

- tr.asteron.example — турецкая маркетинговая платформа

Пользователь спрашивает: «Какие услуги предоставляет Asteron в Турции?» Ответ ИИ: «Asteron Holdings Ltd. самостоятельно организует все туры в Анталии, имеет 400 сотрудников в Турции, и Asteron Club Antalya напрямую отвечает за удовлетворенность клиентов в Анталии.» Реальное положение дел: Материнская компания владеет брендом. Местная компания занимается продажами и заключением контрактов. Компания Mira осуществляет часть операций. Asteron Club Antalya является независимой франшизой. 400 сотрудников — это общее количество сотрудников группы. Франшиза непосредственно отвечает за удовлетворенность клиентов на уровне бизнеса. Ошибки в ответе: Материнская компания представлена как ответственная за предоставление услуг. Общее количество сотрудников группы сообщается так, как если бы это была мощность местной компании. Ответственность франшизы напрямую возлагается на материнскую компанию. Бренд, юридическое лицо и операционный партнер объединены. В ответе бренд мог быть указан правильно. Однако граф сущностей неверен.

Следовательно, оценка GEO не может основываться только на упоминании имени или общей плавности предложения.

22. РЕЗУЛЬТАТЫ РАЗРЕШЕНИЯ СУЩНОСТЕЙ

Для каждого наблюдения результат разрешения сущности можно записать с одним из следующих статусов.

ER-0— НЕ ОЦЕНИВАНО

Разрешение сущности не было оценено.

ER-1— ТОЧНАЯ ЦЕЛЬ

Целевая сущность была разрешена правильно и ясно.

ER-2— ПРИЕМЛЕМОЕ СООТВЕТСТВИЕ НА УРОВНЕ БРЕНДА

Юридическое название не указано, но целевой бренд и корпоративная сущность разрешены правильно в основном.

ER-3— ЧАСТИЧНОЕ ИЛИ НЕДОПОЛНЕННОЕ

Ответ приблизился к правильной области, но не содержит юридической, географической или операционной идентичности.

ER-4— СМЕШЕНИЕ НЕСКОЛЬКИХ СУЩНОСТЕЙ

Несколько отдельных сущностей были объединены в один объект.

ER-5— неверная сущность

Была выбрана другая персона, компания, продукт или доменное имя.

ER-6— ДВОЙСТВЕННОЕ ЗНАЧЕНИЕ

Нельзя определить по ответу, какая сущность была разрешена.

ER-7— ПОДХОДЯЩЕЕ УТОЧНЕНИЕ

Запрос действительно двусмысленный, и система правильно попросила уточнение. Запрос уточнения не обязательно следует считать ошибкой.

23. УСПЕХ В РАЗРЕШЕНИИ СУЩНОСТИ ОТЛИЧАЕТСЯ ОТ ТОЧНОСТИ ПРЕДСТАВЛЕНИЯ

Система может анализировать правильный объект, но при этом предоставлять неверную информацию о нем. Другая система может правильно указать некоторые существенные факты, но связать их с неверная сущность. Существуют четыре основных ситуации:

Анализ сущности	Материальное содержание	Результат
Верно	Верно	Сильный кандидат на правильное представление
Верно	Неверно	Правильный объект, неверный атрибут
Неверно	Частично правильный	Правильная информация, связанная с неверная сущность
Неверно	Неверно	Полный разбор представления

По этой причине анализ идентичности может быть отдельным шлюзом. В частности, ответ с серьезной путаницей с другой сущностью невозможно спасти при высоком среднем показателе фактической точности.

24. СВЯЗЬ РЕШЕНИЯ О СООТВЕТСТВИИ С СУЩНОСТЬЮ

Решение о соответствии должно быть установлено в следующем формате:

> *[ID сущности] + [тип сущности] + [область] + [цифровые каналы] + [страна и языки] + [время] + [версия стандарта]*

Пример шаблона: «Для [бренд/юридическое лицо], идентифицированного с Entity ID NGE-XXXX NOMOS; публичные страницы на турецком и английском языках на указанном домене, в рамках указанного продукта и области услуг, были оценены [дата] в соответствии с [версией стандарта].» Затем должно быть чётко указано: «Это решение не распространяется автоматически на дочерние компании, франшизы, продукты вне области оценки, другие домены и языки, которые будут добавлены в будущем.»

25. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH03-N01

Каждый аудит GEO должен определить уникальный целевой объект до того, как будут видны результаты.

CH03-N02

Нельзя предполагать, что якорь промпта совпадает с целевым объектом.

CH03-N03

Доменные имена, торговые марки, юридические лица, продукты, услуги и личные идентичности должны быть различимы друг от друга.

CH03-N04

Свойство сущности не может быть передано другой связанной сущности без четких доказательств и рамок.

CH03-N05

Лицензия, сертификат, количество сотрудников, клиенты или результаты соответствия материнской компании не могут автоматически применяться к дочерним предприятиям.

CH03-N06

Требование на уровне бренда и сторона, фактически предоставляющая услугу и несущая юридическую ответственность, должны быть разделены.

CH03-N07

Франчайзер, оператор франшизы, дистрибьютор, дилер и продавец на платформе должны быть зарегистрированы как отдельные юридические лица.

CH03-N08

Каждое материальное отношение между сущностями должно содержать информацию о направлении, объеме и времени.

CH03-N09

Отношения, такие как «партнер», «клиент», «эксперт», «группа компаний» и подобные, не могут использоваться неоднозначно.

CH03-N10

Области разрешения коллизий должны использоваться для разных сущностей с одинаковым названием.

CH03-N11

Если неопределенное требование включается в основной рейтинг, метод определения неопределенности должен быть заранее установлен; в противном случае требование должно быть уточнено.

CH03-N12

Объем сущности не может быть расширен или сокращен после получения результатов с целью повышения положительного результата.

CH03-N13

Исторические сущности и отношения не могут использоваться так, как если бы они имели статус текущей сущности.

CH03-N14

Покупка, слияние или передача бренда не автоматически передают прошлые характеристики новому объекту.

CH03-N15

Решение о соответствии может применяться только к единице соответствия, указанной в решении.

CH03-N16

Новый продукт не может автоматически наследовать текущую доступность языка, доменного имени, аффилированной компании или страны.

CH03-N17

Успешность разрешения сущности и фактическая точность должны оцениваться отдельно.

CH03-N18

Объединение материальной идентичности с неверная сущность не может быть сделано невидимым в пределах общего фактического среднего.

CH03-N19

Канонические записи сущностей должны версионироваться и сохранять изменения с течением времени.

CH03-N20

Должен быть ответственный человек или институция за запись сущности.

26. ФОРМЫ ОШИБОК

CH03-F01— РАССМАТРИВАТЬ ДОМЕННОЕ ИМЯ КАК КОМПАНИЮ

Цифровой адрес представлен как юридическая или операционная сущность.

CH03-F02— ОБЪЕДИНЕНИЕ БРЕНДА С ЮРИДИЧЕСКОЙ КОМПАНИЕЙ

Название бренда используется вместо юридического лица, которое заключает контракты и несет ответственность.

CH03-F03— РАССМОТРЕНИЕ ПРОДУКТА КАК КОМПАНИИ

Название продукта определяется как компания или фирменный стиль.

CH03-F04— ПЕРЕНОС ХАРАКТЕРИСТИК ОСНОВАТЕЛЯ В КОМПАНИЮ

Возраст, опыт, награды или предыдущий проект основателя становятся корпоративной характеристикой.

CH03-F05— ПЕРЕНОС ХАРАКТЕРИСТИК ГРУППЫ В МЕСТНУЮ КОМПАНИЮ

Объединенные данные о клиентах, сотрудниках или доходах назначаются местному бизнесу.

CH03-F06— ПЕРЕДАЧА ЛИЦЕНЗИИ ЧЕРЕЗ ОТНОШЕНИЯ

Лицензия материнской компании, продукта, сотрудника или партнера используется в названии другого юридического лица.

CH03-F07— ОБЪЕДИНЕНИЕ ФРАНШИЗЫ И ОСНОВНОГО БРЕНДА

Результаты деятельности локального бизнеса и ответственность представлены напрямую как глобальный бренд.

CH03-F08— ОБЪЕДИНЕНИЕ ПРОИЗВОДИТЕЛЯ И ПРОДАВЦА

Производитель продукта и местный продавец или дистрибьютор рассматриваются как единое целое.

CH03-F09— ОБЪЕДИНЕНИЕ ПЛАТФОРМЫ И ПРОДАВЦА

Платформа маркетплейса считается ответственной за продукт продавца, гарантию или доставку.

CH03-F10— РАССМАТРИВАНИЕ ОТНОШЕНИЙ КАК ПЕРЕДАЧИ ФУНКЦИЙ

Статус клиента, членство, использование технологий или партнёрство ошибочно представляются как передача полномочий.

CH03-F11— ОБНОВЛЕНИЕ ИСТОРИЧЕСКОЙ ИДЕНТИЧНОСТИ

Старое имя, собственность, партнёрство или управляющий используются так, как если бы это была текущая реальность.

CH03-F12— ПРИСВОЕНИЕ ПРОШЛОГО ПОСЛЕ ПОКУПКИ

Все прошлые успехи приобретенного объекта приписываются новому владельцу без объяснения.

CH03-F13— ОБЪЕДИНЕНИЕ ОБЪЕКТОВ С ОДИНАКОВЫМИ НАЗВАНИЯМИ

Разные компании или лица рассматриваются как единый объект из-за сходства названий.

CH03-F14— ИСКУССТВЕННОЕ РАЗДЕЛЕНИЕ ОБЪЕКТА

Неудачный язык, продукт или локальная операция объявляется отдельным объектом для поддержания положительного результата.

CH03-F15— ИСКУССТВЕННОЕ РАСШИРЕНИЕ ОБЪЕКТА

Результаты успешного бренда или продукта распространяются на всю группу.

CH03-F16— УЧЕТ СФЕРЫ ПРИЕМЛЕМОСТИ

Узкий результат аудита используется для всей компании, группы или портфеля продуктов.

CH03-F17— УЧЕТ ЯЗЫКОВОЙ ЮРИСДИКЦИИ

Наличие версии на языке используется как доказательство юридической возможности компании или сервиса в этой стране.

CH03-F18— УЧЁТ ТЕХНИЧЕСКОГО КОНТРОЛЯ КАК СОБСТВЕННОСТИ

Агентство или разработчик показываются как юридический или коммерческий собственник учреждения, потому что они управляют доменным именем.

CH03-F19— ОСТАВЛЕНИЕ ПРИСТАВЛЕННОГО ИСТЦА НЕОПРЕДЕЛЁННЫМ

Невозможно определить, кто несёт ответственность за некорректную информацию о цене, лицензии или вместимости.

CH03-F20— ИНТЕРПРЕТАЦИЯ ЯКОРЯ промпта ПОСЛЕ РЕЗУЛЬТАТА

Целевая сущность, которая сделает ответ успешным, выбирается после этого.

27. ПРОЦЕДУРА АУДИТА

Шаг 1— Запись якоря промпта

Точно какая из следующих характеристик пользователя:

- название,
- доменное имя,
- брендом,
- продукта,
- личность

выражение используется, полностью записывается.

Шаг 2— Идентифицировать целевую сущность

Отмечается, какую сущность на самом деле измеряет аудит. Цель классифицируется как:

- юридическим лицом,

- брендом,
- продукта,
- услуги,
- пакет сущностей

Применимый класс должен быть записан явно.

Шаг 3— Создать канонический идентификатор сущности

Записываются каноническое имя, тип сущности, юридическая идентичность, юрисдикционная зона, псевдонимы, временной статус.

Шаг 4— Построить граф сущностей

Извлекаются направленные отношения между материнской компанией, дочерними компаниями, брендом, продуктом, доменным именем, физическими лицами и партнерами.

Шаг 5— Определить роли ответственности

Разделяются владельцы утверждений, владельцы цифровые каналы, владельцы предоставления услуг, юридически ответственные, владельцы доказательств.

Шаг 6— Проверить риск передачи атрибутов

В ответ:

- лицензия,
- свидетельством,
- сотрудник,
- клиент,
- дата учреждения,
- успех,
- соответствие

была ли она передана на неверная сущность?

Шаг 7— Проверить временную идентичность

Ответ:

- бывшее название,
- бывший владелец,
- прекращенное партнерство,
- бывший сотрудник,
- статус до приобретения

используется ли это?

Шаг 8— Проверка географических и локальных ролей

Правильны ли роли глобального бренда, местной компании, дистрибьютора и франшизы?

Шаг 9— Классификация сущности, разрешенная системой

Назначен соответствующий статус между ER-1 и ER-7.

Шаг 10— Фиксация единицы соответствия

Для применения окончательного решения:

- сущность,
- поле,
- продукта,
- языка,
- страны,
- время

граница указана.

Шаг 11— Ограничение публичного требования

Результат сужается, если его обобщить до учреждения или группы, шире определённой сущности.

Шаг 12— Зафиксировать способ изменения и возражение

Если сообщается об ошибке идентичности или отношения:

- кто будет проверять,
- кто будет изменять,
- какая версия будет сгенерирована

определяется.

28. НЕОБХОДИМЫЕ ДОКАЗАТЕЛЬСТВА

Каноническая идентичность сущности; юридическое название и тип сущности; официальная запись или эквивалентная проверка; право собственности на торговую марку; контроль доменного имени; отношения между материнской и дочерней компаниями; право собственности на продукты и услуги; записи о местном бизнесе и дистрибьюторах; франчайзинговые отношения; договаривающаяся и выставляющая счета сторона; ответственность за предоставление услуг; владелец лицензии и сертификата; основатель и руководящие должности; активные и исторические названия; записи о приобретениях, слияниях и передачах; страна и юрисдикция; официальные языки и записи транслитерации; граф сущностей; претендент; владелец контента; владелец технической интерфейсы; область аудита; объекты и интерфейсы вне области; версия записи сущности; дата последней проверки; и ответственное лицо или организация.

29. ЧЕК-ЛИСТ АУДИТА

Была ли запись якоря запроса выполнена чётко? Была ли целевая сущность идентифицирована до того, как были получены результаты? Были ли юридическое лицо и доменное имя разделены? Были ли бренд и владелец бренда разделены? Были ли продукт и компания разделены? Были ли атрибуты основателя отделены от атрибутов организации? Правильны ли отношения между материнской компанией и дочерней? Отличаются ли франшизы и дистрибьюторы? Были ли идентифицированы стороны, продающие и предоставляющие услугу? Были ли истец и технический издатель разделены? Принадлежат ли лицензии и сертификаты правильной сущности? Указаны ли показатели сотрудников, численности клиентов и доходов в правильном объёме? Были ли исторические названия и

отношения отделены от текущего состояния? Были ли проверены различные сущности с одинаковым названием? Были ли несколько доменов и пользовательских интерфейсов связаны с правильной сущностью?

Связаны ли многоязычные имена с одной и той же канонической идентичностью? Ясны ли роли страны и юрисдикции? Классифицировалась ли сущность отдельно в результате разрешения сущностей? Ясно ли определено подразделение по соответствию? Распространяется ли оно на сущности за пределами сферы решения? Версионируется ли запись сущности и проставлена ли дата? Существуют ли пути для возражений и исправлений при изменениях идентичности?

30. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Пользователь знает торговую марку, а не юридическую компанию».

Правильно. Для пользователя естественно использовать бренд в запросе. Аудит не заставляет пользователя использовать юридические термины. Однако оценка в фоновом режиме:

- бренд,
- владелец,
- поставщик услуг,
- сторона по контракту

должна различать. Пользователю можно дать простой ответ. Журнал аудита не должен быть простым.

Возражение 2— «В малом бизнесе бренд и компания уже являются одним и тем же»

В некоторых небольших структурах:

- один бренд,
- один домен,
- единственный владелец бизнеса,
- один поставщик услуг

Можно встретить. В этом случае граф сущностей простой. Тем не менее, запись должна быть сделана явно. Простота не означает неопределенность.

Возражение 3— «Группа компаний действует как один бренд».

С точки зрения маркетинга может быть один бренд. С точки зрения юридической, операционной и доказательной ответственности могут существовать отдельные сущности. Если решения принимаются на уровне группы, должно быть четко показано, какие функции консолидированы.

Возражение 4— «Столько деталей ненужно усложняет GEO-оценку».

Правильный балл, присвоенный неверная сущность, не полезен. Разрешение сущностей не является дополнительной нагрузкой для аудита, а является условием действительности.

Возражение 5— «Системы ИИ уже не обращают внимания на эти правовые различия.»

Именно по этой причине это должно быть измерено. Система ИИ:

- брендом,
- компания,
- продукта,
- местный бизнес

Пропуск различия между ними является самой ошибкой представления.

Возражение 6— «Если материнская компания контролирует всю группу, характеристики могут быть переданы.»

Контрольные отношения важны в некоторых областях. Однако:

- лицензия,
- контракты,
- клиент,
- сотрудник,
- ответственность за данные,
- соответствие

такие характеристики требуют отдельного подтверждения для переноса. Контроль — это не объединение всех характеристик.

Возражение 7— «Разве достижения старой компании после поглощения не являются частью новой группы?»

Это можно объяснить в историческом контексте. Однако:

- в течение какого периода был успех,
- какой командой и организацией,
- до или после приобретения

Следует указать, когда он был достигнут. Новое владение не меняет историю.

Возражение 8— "промпт уже ясна; пользователь ввел доменное имя."

Доменное имя может быть ясным. Тем не менее: на вопрос "Что представляет доменное имя?" также должен быть дан ответ. Домен может быть компанией, магазином, порталом продукта или многофункциональной платформой.

Возражение 9— "Описывать местного дистрибьютора как бренд проще для пользователя."

Простота не заменяет ответственность. Ответ можно упростить следующим образом: "Бренд предоставляет услуги в Турции через уполномоченного местного дистрибьютора." Это и понятно, и правильно.

Возражение 10— "Все эти отношения меняются со временем."

Это правильно. По этой причине запись сущности:

- датированная,
- версионированная,

- должна быть инициирована изменением

Это не уменьшает необходимость записи вариативности. Наоборот, это увеличивает её.

ОБЩЕЕ ПРАВИЛО ГЛАВЫ 33

Перед генерацией GEO-оценки мы спрашиваем: «Сколько ответов правильные?» Но более фундаментальный вопрос, который следует задать до этого, следующий:

«О чьих именно ответах идет речь?»

Если мы рассматриваем доменное имя как компанию, мы измеряем неправильный объект. Если мы рассматриваем бренд как юридическую компанию, мы неправильно распределяем ответственность. Если мы рассматриваем продукт как компанию, мы искажаем его особенности. Если мы превращаем опыт основателя в историю компании, мы преувеличиваем прошлое. Если мы считаем количество сотрудников группы как местную мощность, мы завышаем ожидания от сервиса. Если материнская компания передает свою лицензию франшизе, мы направляем пользователя к неправильному органу. Если мы приписываем прошлые достижения после приобретения новому владельцу, мы переписываем историческую правду. Если мы расширим аудит узкой доменной области на всю компанию, мы «отбелим» сферу соответствия требованиям. Следовательно, определение сущности — это не просто очистка данных. Это начало представительной ответственности. Система может описывать неверная сущность с правильной информацией. Это всё равно является искажением. Система может описывать правильную сущность с неверной информацией. Это тоже является искажением.

Оба требуются для правильного GEO:

> *Правильная сущность. / Правильный атрибут. / Правильное отношение. / Правильная область. / Правильное время.*

Следовательно, третий закон измерения NOMOS таков:

> *Если измеряемая сущность не ясна, точность измерения также не ясна.*

Четвёртый закон выглядит следующим образом:

> *Отношение не является идентичностью.*

Пятый закон звучит следующим образом:

> *Право собственности не равно передаче всех атрибутов.*

Его шестой закон таков:

> *Соответствие не распространяется на соседние сущности.*

NOMOS Раздел 3 Порядок

> *Не давайте мне только имя. / Покажите, к какой сущности принадлежит это имя.*
>
> *Не рассматривайте домен как компанию, бренд как контрагента или продукт как производителя.*
>
> *Не используйте прошлое основателя как возраст учреждения.*
>
> *Не передавайте лицензию материнской компании аффилированной организации, франшизе или дистрибьютору.*

Не превращайте клиента группы в клиента местного бренда, или награду бренда в награду всех продуктов.

Не описывайте сотрудничество как владение, членство как аккредитацию или использование технологий как партнерство.

Не объединяйте старое название с новым объектом, или приобретенное прошлое с сегодняшними показателями.

Не расширяйте объект, чтобы преувеличить положительные результаты / Не разделяйте объект, чтобы скрыть отрицательные результаты.

Сначала зафиксируйте цель. / Затем зафиксируйте направление каждой связи. / Затем прикрепите каждый атрибут к его истинному владельцу. / Затем добавьте соответствующее время и страну. / Только после этого принимайте решения для этого объекта один.

Заключительное предложение главы

Первый вопрос при проверке GEO - это не то, что говорит ИИ, а точно кто или что касается каждого утверждения.

Нормативное ядро

Каждый аудит GEO ДОЛЖЕН определить и задокументировать версию до того, как будут наблюдаться результаты: - якорь запроса, - целевой объект, - тип объекта, - каноническую запись идентичности, - граф отношений, - временные и юрисдикционные границы, - объект аудита, - и единицу соответствия. Домены, бренды, юридические лица, корпоративные группы, дочерние компании, продукты, услуги, лица, франшизы, дистрибьюторы, платформы и продавцы НЕ ДОЛЖНЫ рассматриваться как взаимозаменяемые объекты. Атрибуты, учетные данные, лицензии, клиенты, сотрудники, показатели, полномочия и соответствие НЕ ДОЛЖНЫ передаваться между связанными объектами, если отношения, область, время и доказательства явно не поддерживают такую передачу. Сущность, обработанная ИИ-продуктом, ДОЛЖНА оцениваться отдельно от фактической точности ответа. Решение о соответствии ДОЛЖНО применяться только к явно идентифицированной сущности и проверяемому объему. Оно НЕ ДОЛЖНО автоматически распространяться на аффилированные лица, продукты, юрисдикции, языки, домены, будущие версии или связанные сущности.

ГЛАВА 04

Два уровня реальности

Официальное представление и проверенная реальность сущности

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K10; K11

Граница главы

Раздел 1 установил следующее положение:

> *Одиночный ответ не является GEO баллом.*

Глава 2 добавила следующее:

> *Представление — это распределение, которое происходит среди пользователей, продуктов ИИ, стран, языков, сессий и времени.*

Глава 3 затем определила объект распределения:

> *Сущность не может быть измерена, пока она не будет определена.*

Теперь мы переходим к следующему обязательному вопросу:

> *По какой записи о фактическом состоянии мы будем оценивать, правильен ли ответ ИИ?*

Какой источник должен определять аудит: сайт аудируемого предприятия, юридическая запись, страница с ценами компании, независимые отчеты, лицензирующий орган, отзывы клиентов, собственные данные о производительности компании или утверждение конкурента? Когда источники противоречат друг другу, какой из них следует принимать за правильный? Если компания публикует ложное или неподтвержденное утверждение, а ИИ повторяет его идеально, значит ли это, что система достигла успеха? Если официальный сайт устарел, но ИИ дает другой ответ на основе актуальных проверенных внешних данных, значит ли это, что система провалилась? Пока эти вопросы не будут решены:

- правильное представление,
- неправильное представление,
- верность источника,
- фактическая точность,
- реальность сущности

Эти термины иногда используются так, как будто они относятся к одной и той же концепции. Эта глава устраняет возникающую путаницу, различая два уровня реальности:

Первый уровень

Слой официального представления

Показывает, что сущность говорит о себе на канонических интерфейсах, которыми она управляет.

Второй уровень

Проверенный слой реальности сущности

На основании имеющихся доказательств показывается, какие материальные утверждения поддерживаются в каком объеме и за какой период времени. Эти два слоя взаимосвязаны. Но они не совпадают. Первый слой отвечает на вопрос: «Как субъект определяет себя?» Второй слой отвечает на вопрос: «Какая часть этого определения поддерживается какими доказательствами?» Первая книга определяет правильный стандарт, вторая книга разделяет формы нарушений, а система аудита предназначена для измерения и доказательства этих различий. Этот раздел выполняет задачу аудита на уровне утверждений.

NOMOS Вызов

Веб-сайт вашей компании содержит следующее предложение: «Мы — ведущая в мире GEO-компания». Продукт ИИ читает эту страницу и отвечает пользователю: «Эта компания является ведущей в мире GEO-компанией». Правильно ли ответ отражает информацию на вашем сайте? Да. Правильно ли ответ отражает проверенную реальность?

Мы этого ещё не знаем. Может быть:

- существуют сравнительные исследования,
- масштаб рынка был определён,
- Критерий лидерства был опубликован,

результат был независимо проверен. Возможно, ничего из этого не существует. Если бы ИИ передал собственное утверждение компании: «Компания определяет себя как мирового лидера среди GEO-компаний.», он сохранил бы источник утверждения. Однако, когда он сказал: «Это мировая ведущая GEO-компания.», компания могла приподнять это утверждение до независимой реальности.

Рассмотрим другой случай. На официальном сайте компании указано: «Мы предоставляем услуги только в Турции». Тем не менее, проверенные данные показывают, что компания начала работать в Германии и Великобритании шесть месяцев назад, а сайт не был обновлен. Исходя из текущих юридических и операционных данных, ИИ отвечает: «Компания предоставляет услуги в Турции, Германии и Великобритании».

ИИ не полностью соответствует официальному сайту. Тем не менее, может ли он быть ближе к проверенной текущей реальности? Да. Рассмотрим третий случай. На сайте компании указано: «Мы работаем в 20 странах». Независимые и внутренние записи подтверждают только восемь. ИИ отвечает: «Компания работает в 20 странах».

Этот ответ:

- верный официальному представлению,
- вопреки проверенной реальности

Это возможно. Первая статья этого раздела следующая:

> *Верность официальной записи не то же самое, что фактическая точность.*

Второе положение гласит:

> *То, что компания говорит, и то, что можно доказать, следует записывать отдельно.*

Её третье положение гласит:

> *Искусственный интеллект, который правильно повторяет официальную ошибку, не является точным представлением.*

Четвертая статья гласит:

> *Искусственный интеллект, который расходится с устаревшей официальной записью, не является автоматически неправильным, если у него есть надежные доказательства.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — установить два отдельных справочных уровня, которые будут использоваться при оценке GEO-1000. Раздел делает следующие нормативные различия:

- Заявление о существовании против доказанного факта
- Официальный источник против независимого источника
- Источник первой стороны против независимой проверки
- Верность источнику против материальной точности
- Повторение утверждения против проверки утверждения
- Текущая официальная запись против старой официальной записи
- Официальное молчание против внешне подтвержденной информации
- Перспектива компании против внешней оценки реальности
- Официальное доказательство против официального утверждения
- Независимость источника против релевантности источника по отношению к теме
- Официальная ошибка против ошибки ИИ
- Конфликт ссылок против сбоя модели
- Неопределенность против сбоя
- Интерес истца против истины заявления
- Наличие доказательств и то, поддерживают ли доказательства заявленное утверждение

К концу этого раздела для каждого ответа ИИ должны быть заданы два отдельных вопроса:

> *Верно ли ответ передает официальное и каноническое представление о сущности?*

и:

> *Правильно ли ответ отражает материальную реальность, подтверждаемую в настоящее время проверяемыми доказательствами?*

Эти вопросы могут дать одинаковый ответ. Они также могут дать разные ответы. Когда они различаются, роль аудита не в том, чтобы скрыть разницу, а в том, чтобы ее объяснить.

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

> *Каждый существенный ответ ИИ следует оценивать на основе двух отдельных ссылок без преобразования: Записи официального представления и Записи проверенной реальности сущности.*

Ответ:

- верный официальному представлению,

- соответствующее проверенной реальности,
- соответствующий обеим,
- далёкий от обеих

Это возможно. NOMOS не может тихо объединить эти два результата в одно решение «истина–ложь». Высокая производительность на первом уровне не компенсирует неудачу на втором уровне. В частности:

> *Ответ III, который правильно повторяет непроверенное или ложное утверждение, опубликованное самой сущностью, может демонстрировать соблюдение официального представления; однако он не может демонстрировать точность проверенной реальности сущности.*

3. ПЕРВЫЙ УРОВЕНЬ

СЛОЙ ОФИЦИАЛЬНОГО ПРЕДСТАВЛЕНИЯ

Слой официального представления — это версионированная запись текущих материальных утверждений, которые он публикует о себе на канонических ресурсах, контролируемых или явно авторизованных организацией. Этот слой отвечает на вопрос:

> *Как субъект хочет, чтобы его понимали, и какую информацию он официально публикует?*

Этот слой включает в себя:

- заявление о юридической идентичности,
- отношение бренда к компании,
- определения продуктов и услуг,
- публикуемые цены,
- страны обслуживания,
- условия поддержки,
- официальные политики,
- публично раскрытые границы,
- Он может содержать официальную информацию о валюте и версиях

Слой официального представления важен, потому что система ИИ должна правильно понимать, как организация:

- определяет себя,
- какие услуги он предлагает,
- какую цену публикует,
- какие условия принимает

Однако слой официального представления не является автоматическим независимым доказательством каждой информации. Заявление на собственном веб-сайте учреждения, особенно в отношении требований к внешнему миру, таких как превосходство и производительность, представляет само утверждение; оно само по себе не доказывает его правильность.

4. ЧТО ТАКОЕ ОФИЦИАЛЬНЫЙ ИСТОЧНИК?

Источник не считается официальным только потому, что на нем есть логотип учреждения. Официальный источник должен обладать как минимум соответствующими признаками следующего характера:

- Находится под контролем организации
- Проходит авторизованный процесс публикации
- Одобрен владельцем фактической информации
- Канонический URL или идентификатор записи
- Версия
- Дата публикации или проверки
- История изменений
- Ответственное лицо или организация
- Область
- Статус языка и локали
- Архивный или актуальный статус

Мнение, опубликованное на личной странице в социальных сетях сотрудника, не является автоматически официальным заявлением компании. Черновик файла агентства не является официальной публикацией. Электронное письмо от торгового представителя, если оно не уполномочено, не является общей корпоративной политикой. Старый PDF на сайте компании может не заменять новый официальный документ.

5. ПОЯВЛЕНИЕ ОФИЦИАЛЬНОГО ПРЕДСТАВЛЕНИЯ

Согласно соответствующему аудиту, официальные источники могут включать:

- Основной корпоративный веб-сайт
- Юридические оговорки
- Страницы продуктов и услуг
- Страницы с ценами и коммерческими предложениями
- Официальный справочный центр
- Документы по связям с инвесторами
- Публичные политики и контракты
- Официальные пресс-релизы
- Структурированные данные
- Канонические записи JSON
- Официальные выводы API
- Официальный контент в приложении
- Авторизованные локальные сайты
- Официальные аккаунты в социальных сетях
- Заявления, поданные в публичные учреждения
- Записи о выпуске и отзыве

Эти интерфейсы могут противоречить друг другу. В этом случае «официальное представление» не является единой, чистой реальностью. Это набор записей, которые должны проверяться внутренне.

6. ЗАПИСЬ ОБ ОФИЦИАЛЬНОМ представлении

Для каждого проверяемого объекта должна быть создана запись об официальном представлении. Эта запись не является копией всего официального контента. Это прослеживаемое резюме фактических утверждений и границ. Каждая запись должна включать следующие поля в мере возможности:

- Идентификатор утверждения
- Полный текст утверждения
- Объект утверждения
- Тип утверждения
- Канонический источник
- Владелец источника
- Язык и локаль
- Дата публикации
- Дата последней проверки
- Срок действия
- Официальный статус
- Область
- Ограничения
- Предыдущая версия
- Причина изменения
- Истец
- Одобрение человеком

Без записи об официальном представлении объекта предложение «ИИ достоверно отражает сайт» не может быть однозначно проверено.

7. АВТОРИТЕТ ОФИЦИАЛЬНОГО ИСТОЧНИКА ЗАВИСИТ ОТ ТИПА УТВЕРЖДЕНИЯ

Не все официальные источники имеют одинаковый вес для каждого утверждения. Авторитет источника следует оценивать в соответствии с типом утверждения.

7.1. Решения в рамках собственного контроля организации

Пример:

- Какие услуги она предоставляет общественности
- Опубликованная цена
- Часы поддержки
- Официальная политика возврата
- Канал связи

- Общедоступная версия продукта

В этих областях официальный источник является фундаментальным и часто основным авторизованным документом. Однако, если официальная страница противоречит фактическому процессу, одной только страницы недостаточно. Например:

- страница с ценой 1,000 евро,
- оформление заказа 1,500 евро

Если это отображается, существует противоречие в слое официального представления.

7.2. Юридическая личность и полномочия

Пример:

- Регистрация компании
- Лицензия
- Профессиональные полномочия
- Право собственности на товарные знаки
- Налоговый или юридический статус

Веб-сайт компании может раскрыть эту информацию. Однако официальный документ или авторизованное учреждение могут предоставить более надежное подтверждение.

7.3. Заявления о результативности и успехе

Пример:

- «98% наших проектов успешны.»
- «Мы увеличиваем доход наших клиентов в среднем на 40%.»
- «У нас самый высокий уровень удовлетворённости.»

Веб-сайт компании может быть владельцем заявления. Но чтобы поддержать заявление:

- data,
- метод,
- знаменатель,
- период,
- неудачные результаты,
- независимый обзор, если возможно

требуются.

7.4. Заявления о превосходстве и лидерстве

Пример:

- "Мировой лидер"
- "Лучший"
- "Номер один"
- "Единственный настоящий стандарт"
- "Самая надежная компания в отрасли"

Заявление сайта не является независимым сравнением. Такие утверждения:

- требуют вселенной для сравнения,
- критериев,
- метод,

- дата,
- объем охвата,
- независимости

Оба условия должны быть установлены.

7.5. Отношения с третьими сторонами

Пример:

- Клиент
- Партнер
- Спонсор
- Членство
- Аккредитация
- Технологическое партнерство
- Дистрибьюторство

Одностороннее официальное заявление может быть ценным. Однако оно может потребовать проверки другой стороной отношений или надежным документом.

7.6. Результаты пользователя

Пример:

- Удовлетворенность клиентов
- Фактический результат использования
- Уровень возврата
- Уровень жалоб
- Повторная покупка

Данные компании в этих областях являются важными доказательствами первой стороны. Однако:

- выбор данных,
- неявка,
- методология,
- конфликт интересов

должны быть раскрыты.

8. Точность официального представления

Понятие, указывающее, в какой степени ответ ИИ соответствует слою официального представления:

Точность официального представления

Канонический английский термин:

Точность официального представления

Это измерение задает следующие вопросы: Правильно ли ИИ интерпретирует сущность? Точно ли оно передает официальные продукты и услуги? Искажает ли оно официальную цену? Убирает ли официальные границы? Расширяет ли официальное покрытие страны? Представляет ли мнение компании как независимый факт? Использует ли устаревшую

официальную информацию как актуальную? Скрывает ли противоречия между официальными записями? Добавляет ли материальные утверждения, отсутствующие в официальном источнике? Сохраняет ли качество источника? Точность официального представления — это измерение того, насколько точно сущность передает реальность, которую она опубликовала. Это измерение:

> *Это показывает не то, что утверждения сущности верны, а то, в какой степени IIII точно передаёт то, что говорит сущность.*

9. ВТОРОЙ СЛОЙ

ПРОВЕРЕННЫЙ СЛОЙ РЕАЛЬНОСТИ СУБЪЕКТА

Слой реалий проверенной сущности представляет собой версионированную запись оценки материальной информации, существовавшей на конкретную дату измерения, и поддержанную доказательствами, соответствующими типу утверждения. Этот слой отвечает на вопрос:

> *Какие утверждения о сущности действительно подтверждаются какими доказательствами?*

Этот слой включает в себя:

- записи, поступающие от самой компании,
- официальные публичные записи
- независимые источники,
- записи о контрактах или сделках,
- лицензионные документы,
- проверенные исторические записи,
- контрдоказательства,
- неблагоприятные выводы,
- неопределенности,
- остающиеся споры

могут оцениваться вместе.

10. «ПРОВЕРЕННАЯ РЕАЛЬНОСТЬ» НЕ ЯВЛЯЕТСЯ АБСОЛЮТНОЙ РЕАЛЬНОСТЬЮ

Этот протокол не утверждает: «Мы с уверенностью знаем всю абсолютную правду об учреждении.» Проверенная реальность сущности заключается в следующем:

- основанной на доказательствах,
- зависящей от времени,
- зависящей от области применения,
- подверженной изменениям при появлении новых доказательств,
- может быть оспорена,
- может быть отозвана,

может содержать неопределенность. Правильное определение:

> *В указанные дату и область применения, наиболее сильная и проверяемая запись сущности, поддерживаемая доступными доказательствами.*

Следовательно, запись проверенной реальности:

- версионированная,
- с источниками,
- подлежащей возражению,
- поддающейся изменениям,
- осведомленной о неопределенности.

NOMOS не может позиционировать себя как обладателя истины. Он лишь классифицирует, какое утверждение допускают доказательства и в какой степени.

11. ЗАПИСЬ ПРОВЕРЕННОЙ РЕАЛЬНОСТИ СУЩНОСТИ

Полная техническая структура этого слоя будет подробно изложена в одной из последующих глав книги. На базовом уровне для каждого фактического утверждения должны существовать следующие поля:

- Идентификатор утверждения
- Субъект сущности
- Связь или атрибут
- Значение
- Область
- Страна или юрисдикция
- Время
- Источники
- Типы источников
- Проверка источника
- Независимость
- контрдоказательства
- Достаточность доказательств
- Статус проверки
- Неопределенность
- Решение судебного органа
- Статус апелляции
- Дата последней проверки
- Версия записи

Эта запись не может быть создана исключительно на основе положительных доказательств. Записи, которые сужают реальность или противоречат утверждению, также должны сохраняться.

12. СТАТУСЫ ВЕРИФИКАЦИИ

Каждое материальное утверждение может быть классифицировано с одним из следующих статусов:

VR-0— НЕ ОЦЕНИВАЛОСЬ

Утверждение ещё не было оценено.

VR-1— ТОЛЬКО ОФИЦИАЛЬНОЕ УТВЕРЖДЕНИЕ

Утверждение основано исключительно на официальном заявлении самого учреждения. Оно не было опровергнуто. Независимой или достаточной дополнительной проверки нет.

VR-2— ПОДТВЕРЖДЕННЫЕ ДОКАЗАТЕЛЬСТВА ПЕРВОЙ СТОРОНЫ

Утверждение подтверждается следующими данными учреждения:

- транзакцией,
- операцией,
- контракты,
- data,
- записью

подтверждено доказательствами. Источник — первая сторона.

VR-3— ПОДТВЕРЖДЕНО

Утверждение подкреплено множеством соответствующих и согласованных данных. Не все источники должны быть независимыми.

VR-4— НЕЗАВИСИМО ПРОВЕРЕНО

Утверждение подкреплено соответствующим независимым источником или обзором по теме. Независимость сама по себе не означает достаточность. Также требуется, чтобы источник был релевантен теме.

VR-5— ПОДДЕРЖАНО МНОЖЕСТВОМ ИСТОЧНИКОВ

Утверждение сильно подтверждается надежными источниками различных типов в пределах определенного времени и объема.

VR-C — ОПРОВЕРЖЕНО

Утверждение противоречит более сильным или подходящим доказательствам.

VR-U — НЕ ОПРЕДЕЛЕНО

Доказательства недостаточны или противоречивы. Невозможно принять надежное решение.

VR-T — ВРЕМЕННО ОГРАНИЧЕНО

Утверждение подтверждается только за указанный период.

VR-S — ОГРАНИЧЕННЫЙ ПО ОБЛАСТИ ДЕЙСТВИЯ

Утверждение подтверждается только для указанного продукта, страны, пользователя или области применения. Утверждение может иметь несколько статусов одновременно. Пример:

VR-4+ VR-T + VR-S

Это может указывать на то, что утверждение было независимо проверено; однако оно имеет ограничения по времени и объему.

13. НЕЗАВИСИМОСТЬ НЕ ТОЖДЕСТВЕНА ПРАВИЛЬНОСТИ

Независимый источник не всегда является лучшим источником. Блогер может оценивать текущую цену компании, в то время как официальная страница компании с ценами может быть более прямым и уместным источником. Наоборот, когда компания заявляет на своем собственном сайте, что она «лидер отрасли», независимое исследование с раскрытой методикой может быть более подходящим доказательством. Следовательно, оценка источника предполагает больше, чем просто выяснение, является ли источник независимым. Является ли он авторитетным для этого типа утверждения? Насколько близко он к самому утверждению? Соответствует ли его охват и актуальность ситуации? Кто им управляет? Раскрыта ли его методика? Рассматривает ли он контрдоказательства? Может ли он быть проверен? Существует ли конфликт интересов? Самостоятельность сама по себе не гарантирует точность.

То, что источник является первоисточником, не делает его бесполезным. Важно следующее:

> *Это наличие соответствующих полномочий и квалификации доказательств для утверждения, представленной источником.*

14. КАРТА ПРИГОДНОСТИ ИСТОЧНИКА ИСКА

Ниже представлена общая базовая таблица. Конкретные требования к доказательствам могут различаться в зависимости от сектора и уровня риска.

Тип утверждения	Основной источник или доказательство	Роль официального сайта	Необходимость дополнительной проверки
Юридическое название компании	Уполномоченная официальная запись	Показывает декларацию	Обычно требуется
Право собственности на товарные знаки	Регистрация торговой марки, контракт, корпоративная запись	Показывает декларацию	Требуется в случае финансовой информации
Текущая опубликованная цена	Каноническая цена/запись о покупке	Обычно основной	Проверяем с помощью рабочего процесса
Объем услуги	Операционные записи, контракты, официальная страница	Основное заявление	Дополнительные доказательства фактической поставки
Лицензия	Лицензирующее учреждение или реестр	Показывает декларацию	Требуется
Взаимоотношения с клиентом	Контракт, одобрение клиента, двусторонняя запись	Может быть односторонним утверждением	Требуется в случае финансовой информации
Уровень успеха	Исходные данные, метод, знаменатель, аудит	Публикует утверждение	Категорически требуется
Руководство	Сравнительный метод и данные	Самооценка	Необходима независимая или сильная сравнительная оценка
Удовлетворенность клиентов	Репрезентативное исследование, метод, неучтённые ответы	Может объяснить результат	Необходима методология
Дата учреждения	Официальные записи и исторические документы	Показывает декларацию	Правовая и брендовая история должны быть разделены
Партнёрство	Билатеральное заявление или контракт	Одностороннее заявление	Обычно требуется

Тип утверждения	Основной источник или доказательство	Роль официального сайта	Необходимость дополнительной проверки
Сертификат	Учреждение, выдающее сертификат, и область применения	Показывает декларацию	Требуется
Активный сотрудник	Запись в отделе кадров или контракт, публичный профиль	Может показать заявление	Роль и проверка даты
Результат пользователя	Запись о транзакции и результате	Описывает случай	Контекст, метод и требуемое разрешение
"Лучший вариант"	Определённый пользователь и критерии сравнения	Маркетинговый взгляд	Требуется сильное сравнение

Эта таблица не является универсальным рейтингом источников. Она показывает соответствие между утверждением и доказательством.

15. МАТРИЦА ДВОЙНОГО СЛОЯ УТВЕРЖДЕНИЙ

Для каждого финансового требования задаются два отдельных вопроса: было ли оно официально опубликовано? Каков его статус в проверенном реестре фактов? Эти два вопроса составляют следующие основные ситуации.

15.1. Официально и подтверждено

Утверждение: оно присутствует в официальной записи и подтверждается соответствующими доказательствами. Пример: Цена на официальной странице с ценами также проверяется в процессе транзакции. Это один из самых сильных типов основной представления. Если ИИ передаёт утверждение с правильным охватом и временем, успех достигается на обоих уровнях.

15.2. Официально, но только самозаявленный

Утверждение является официальным. Однако дополнительной проверки нет. Пример: «Мы — самая инновационная компания в отрасли». ИИ может передать утверждение так: «Компания описывает себя как самую инновационную в отрасли». Он не может повышать его до: «Это самая инновационная компания в отрасли».

15.3. Официально, но противоречиво

Утверждение содержится в официальных документах. Оно противоречит более подходящим или актуальным данным. Пример: на сайте указано, что 20 страны. Операционные записи показывают восемь стран. Если ИИ точно повторяет сайт, официальная достоверность может быть высокой. Проверенная точность низкая. Эта ситуация:

Точное воспроизведение официальной ошибки

может быть классифицировано как.

15.4. Неофициально, но проверено

Утверждение не найдено на официальном сайте. Однако оно подтверждено другими надежными источниками. Пример: недавнее изменение юридической собственности еще не внесено на официальный сайт. Оно отображается в авторизованном реестре. ИИ может точно передавать эту информацию с указанием правильного источника и времени. Официальное представление может иметь низкую достоверность или отличаться. Проверенная фактическая точность может быть высокой.

15.5. Неофициально и непроверено

Утверждение: отсутствует в официальной документации, недостаточно внешних доказательств. ИИ не должен передавать это так, как будто это правда.

15.6. Лимит материала официально пропущен

Компания публикует положительное утверждение на своей официальной странице. Однако она не публикует необходимые ограничения для его правильного понимания. Пример: «Мы предоставляем услуги в Европе». Однако услуга предлагается только в двух странах и удаленно. Если ИИ просто повторяет общее предложение, он может оставаться верным официальному тексту. Однако это может привести к существенному введению в заблуждение. Поэтому Слой официального представления должен фиксировать не только опубликованные предложения, но и существенные упущения.

15.7. Спор не разрешен

Официальные и внешние записи противоречат друг другу. Ни одна из них недостаточно надежна. Правильный результат:

НЕРАЗРЕШЕНО

не должно быть. От ИИ нельзя ожидать вынесения окончательного суждения.

16. ТАБЛИЦА ДВОЙНОЙ ОЦЕНКИ

Верность официальному представлению	Проверенная точность реальности	Базовый комментарий
Высокая	Высокая	Верное и проверенное представление
Высокая	Низкая	Точное воспроизведение официальной ошибки
Низкая	Высокая	Корректив ное представление, которое расходится с официальной записью, но подтверждается доказательствами
Низкая	Низкая	Общее несоответствие представления
Неясно	Высокая	Официальный источник противоречив; внешняя проверка сильна
Высокая	Неясно	Официальное заявление правильно передано; проверка реальности недостаточна
Неясно	Неясно	Недостаточная ссылка для принятия решения

Эта таблица закрепляет важный принцип:

> *Несоответствие ІІІ официальной записи не всегда является ошибкой.*

Точно так же:

> *Соответствие ІІІ официальной записи не всегда является успехом.*

17. ПРАВИЛО СОХРАНЕНИЯ ИСХОДНОГО СТАТУСА

Ответ ИИ должен сохранять не только содержание, но и эпистемический статус содержания. Следующие предложения не одинаковы:

- "Компания является ведущей в мире организацией GEO."
- "Компания называет себя ведущей в мире организацией GEO."
- "Компания заявляет, что является ведущей мировой географической организацией на основе опубликованных исследований, которые еще не были независимо подтверждены."
- "В указанном независимом сравнении компания заняла первое место."

Каждое предложение несет разный статус доказательства. ИИ:

- самостоятельное заявление как независимый результат,
- маркетинговое заявление как научное открытие,
- пресс-релиз как редакционная новость
- спонсируемая награда как независимая аккредитация,
- данные компании как внешний аудит

Это обращение запрещено следующим правилом:

Правило защиты статуса цитирования и источника

Статус источника должен оставаться явным.

18. ЭПИСТЕМИЧЕСКОЕ ПОВЫШЕНИЕ

Укрепление статуса источника утверждения в ответе ИИ называется:

Эпистемическое повышение

Примеры преобразований:

Исходный текст	Неправильное повышение статуса ИИ
«Согласно собственным данным компании»	«Доказано»
«Сообщено конкретными клиентами»	«Работает у всех клиентов»
«Наблюдалось в пилотном исследовании»	«В целом эффективно»
«Компания определяет себя как лидера»	«Лидер отрасли»
«Независимая проверка не проводилась»	Ограничение полностью снято
«Доступно на некоторых рынках»	«Доступно по всему миру»

Эти преобразования не являются резюме. Они производят новые и более сильные утверждения.

19. ОФИЦИАЛЬНОЕ МОЛЧАНИЕ

Отсутствие информации на официальной странице компании может не означать:

- что информация неверна,
- что информация запрещена,

ИИ не может использовать эту информацию. Официальное молчание может принимать три разные формы.

19.1. Немасштабное молчание

Информация может быть за пределами запроса и решения пользователя. Использовать её или нет — не имеет значения.

19.2. Законная дополнительная внешняя информация

Это может быть информация, не найденная на официальном сайте, но подтвержденная надежным источником. Пример:

- Юридическая запись
- Историческое событие
- Публичное нормативное решение
- Независимое исследование

ИИ может использовать эту информацию при соответствующем контексте и статусе источника.

19.3. Молчание о материальном ограничении

Субъект может не раскрывать предел, который влияет на решение пользователя, в своем официальном представлении. Пример:

- Лицензия действительна только в некоторых странах
- Цена, по всей видимости, не включает дополнительную плату
- Услуга доступна только для определенных типов клиентов
- Данные об успехе поступают только из выбранных случаев

В этом случае само официальное представление может содержать вывод аудита. Повторение AI молчания может снизить правдивость реальности.

20. СУЩЕСТВЕННОЕ НЕДОСТАТКИ В ОФИЦИАЛЬНОМ ПРЕДСТАВЛЕНИИ

Официальное предложение может быть не напрямую ложным. Однако удаление необходимых ограничений может сделать его вводящим в заблуждение. Пример: «Мы работаем в Европе». Существенный контекст:

- только в двух странах,
- только удаленно,
- без местного офиса,
- только для корпоративных клиентов

может быть предоставлен. Слой официального представления должен записывать эти две области вместе:

- Основное опубликованное утверждение
- Необходимые материальные ограничения для правильного понимания утверждения

Если ответ ИИ содержит основное предложение, но убирает ограничения:

- верно с точки зрения слов,
- вводящий в заблуждение с точки зрения суждения

может быть.

21. НЕПРАВИЛЬНО — ЭТО НЕПРАВИЛЬНО, ДАЖЕ ЕСЛИ ЭТО ВЫГОДНО ВАМ

Когда продукт ИИ изображает субъект как более могущественный, крупный или авторитетный, чем он есть на самом деле, молчание организации становится проблемой измерения и управления. Местная компания может быть представлена как глобальный бизнес; консультант как регулятор; самопроверка как независимая сертификация; или деятельность, охватывающая восемь стран, как работающая в 20. Даже когда ошибка выгодна организации, она должна быть зафиксирована как ошибка в слое проверенной реальности. Основное предупреждение второй книги устанавливает ту же границу: коммерчески привлекательная ошибка всё равно должна быть исправлена. Когда организация сознательно использует выгодную ошибку ИИ:

- обязанность исправлять
- целостность представления
- коммерческое обещание
- управление

также можно рассматривать с точки зрения

22. КОНФЛИКТ ССЫЛОК

Иногда проблема заключается не в ответе ИИ, а в справочных данных. Пример: на главной странице показывается одна цена. В PDF-файле указана другая цена. При оформлении заказа отображается еще одна цена. Структурированные данные содержат устаревшую цену. В этом случае может не существовать единственного правильного официального представления. Другой пример: в юридическом реестре указана одна форма собственности. На сайте компании указана другая форма собственности. Даты обновления неизвестны. В этом случае проверенная запись о реальности также может содержать расхождение.

23. ПРАВИЛО ДЕФЕКТА ССЫЛКИ

> *Если справочные записи существенно противоречивы или недостаточны, эту неопределенность нельзя автоматически превращать в сбой продукта ІІІ.*

Правильные статусы:

REFERENCE_CONFLICT

НЕДОСТАТОЧНО ДОКАЗАТЕЛЬСТВ

НЕРАЗРЕШЕНО

ВРЕМЕННОЕ РАЗДЕЛЕНИЕ

РАЗДЕЛЕНИЕ ОБЛАСТИ

возможны. Аудит должен разделять вопросы: «Дал ли ИИ неверный ответ?» и: «Была ли необходимая реальная информация, чтобы ИИ дал правильный ответ?» Если учреждение оставило свои официальные записи противоречивыми, мы можем измерить, какую запись

выбирает модель. Однако мы не можем скрывать недостаток исходных данных, ожидая единственного и окончательного институционального ответа от модели.

24. БЛОК КАНОНИЧЕСКОЙ ЗАПИСИ

Оба справочных слоя GEO-1000 должны быть версиями до начала волны измерений:

Официальная запись представителя

Какие страницы? Какие языковые версии? Какая дата? Какие утверждения? Каков статус архива?

Проверенная запись реальности сущности

Какие доказательства? Какой контрдоказательства? Какие статусы проверки? Какие возражения? Какая дата? Если во время измерения происходит существенное изменение: волну можно остановить. Запись можно разделить на новую версию. Предыдущие и последующие наблюдения можно анализировать отдельно. Смешивание результатов под одной ссылкой запрещено.

25. БРЕМЯ ДОКАЗАНИЯ В ЗАВИСИМОСТИ ОТ СИЛЫ УТВЕРЖДЕНИЯ

По мере роста утверждения, бремя доказательства также увеличивается. Представленная ниже лестница является конструктивным изображением.

C0 — Простое утверждение идентичности

"X — это консалтинговая компания." Требуется правильная запись о субъекте и деятельности.

C1 — Утверждение о сфере деятельности

"X предоставляет услуги в Германии." Требуется запись о деятельности, контракте или поставке.

C2 — Количественное утверждение о результатах

"90% проектов выполнено вовремя." Требуются данные о знаменателе, периоде, методе и неудачных результатах.

C3 — Сравнительное утверждение

«Это быстрее, чем у конкурентов». Необходимы вселенная сравнения и эквивалентное измерение.

C4 — Утверждение о превосходстве

«Это лидер отрасли». Требуются отрасль, критерии, география, время и метод.

C5 — Гарантия или универсальное положение

«Гарантирует результат для каждого клиента.» Истинная проверка требует чрезвычайно сильных, всеобъемлющих и применимых доказательств. Официальные источники могут быть надёжными для утверждений C0 и некоторых C1. Сами по себе они могут быть недостаточны на уровнях C3–C5.

26. ДВЕ ОТДЕЛЬНЫЕ МЕРЫ ТОЧНОСТИ

Две основные меры должны быть указаны отдельно в GEO-1000.

26.1. Точность официального представления

Отвечает на следующий вопрос:

> *В какой мере ответ ИИ точно, полностью и с сохранением статуса источника передает версионированное официальное и каноническое представление сущности?*

Это измерение:

- официальная идентичность,
- официальная сфера обслуживания,
- Опубликованная цена,
- официальные границы,
- официальный характер источника

Оценка их областей.

26.2. Точность проверенной реальности сущности

Отвечает на следующий вопрос:

> *В какой степени ответ ИИ точно, всесторонне и с сохранением границ передает текущую реальность сущности, подкрепленную имеющимися доказательствами?*

Это измерение:

- Поддерживаемые фактические утверждения,
- Противоречивые утверждения,
- Устаревшая информация,
- Неправильные связи,
- Фактические недостатки,
- эпистемические улучшения

оцениваются.

27. НЕСПОСОБНОСТЬ ДВУХ ОЦЕНОК КОМПЕНСИРОВАТЬ ДРУГ ДРУГА

Высокая официальная достоверность не может компенсировать низкую проверенную точность. Пример: официальный сайт публикует ложное утверждение. ИИ повторяет это утверждение точно. Официальная достоверность может быть высокой. Проверенная точность низкая. Аналогично, высокая проверенная точность не полностью исключает искажение официального представления. Пример: ИИ правильно использует актуальный внешний источник. Он путает описание продуктов и услуг одной компании с другой. Некоторые внешние факты в ответе могут быть верными. Достоверность официальной идентичности слабая. Поэтому два результата должны появляться отдельно.

28. ДВУХСЛОЙНЫЕ СТАТУСЫ ОТВЕТА

Ответ ИИ может получить один из следующих результатов:

ДВОЙНОЙ ПРОХОД

Это материально согласуется как с официальным представлением, так и с проверенной реальностью.

ОФИЦИАЛЬНО-ПРОХОД / ПРОВЕРЕНО-НЕУДАЧА

Верно передает официальную запись. Однако передает неподтвержденное, устаревшее или противоречивое утверждение официальной записи так, как будто оно является правдой.

ОФИЦИАЛЬНО-НЕУДАЧА / ПРОВЕРЕНО-ПРОХОД

Не полностью соответствует официальному сайту. Тем не менее, он правильно передает истину, подкрепленную актуальными и более убедительными доказательствами. Желательно объяснение различий в ответе.

ДВОЙНАЯ ОШИБКА

Он существенно нарушает как официальную запись, так и проверенную реальность.

ОФИЦИАЛЬНО-СООТВЕТСТВУЕТ / ПРОВЕРЕНО-НЕИЗВЕСТНО

Официальное заявление было точно передано. Статус проверки реальности этого утверждения еще не определен.

КОНФЛИКТ-ССЫЛОК

Справочные записи противоречивы сами по себе. Невозможно дать окончательную оценку ответу.

ПОДХОДЯЩАЯ-НЕОПРЕДЕЛЕННОСТЬ

ИИ правильно объяснил неопределённость ссылки и избегал вынесения окончательного суждения. В некоторых случаях это может быть самым точным ответом.

29. СИНТЕТИЧЕСКИЙ APPLE.COM ДИСПЛЕЙ

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Следующие утверждения, ответы и оценки являются полностью вымышленными. Они не представляют реальную информацию, показатели или ответы какого-либо фактического продукта Apple Inc. Искусственная официальная запись представления: O-01: Компания определяет себя как глобальный бренд потребительских технологий. O-02: Компания публикует определенные категории продуктов и цифровых услуг. O-03: Компания определяет себя как 'самую инновационную компанию в мире'. O-04: Официальный веб-сайт содержит условия магазинов для конкретных стран. Синтетическая проверенная запись реальности: V-01: Корпоративная идентичность и основная категория бизнеса поддерживаются. V-02: Определенные категории продуктов поддерживаются. V-03: Не существует установленного метода сравнения для утверждения 'самая инновационная компания в мире'. V-04: Условия магазинов различаются в зависимости от страны.

29.1. Синтетический ответ A

«Apple.com— это один из официальных корпоративных интерфейсов, где компания представляет свои потребительские технологические продукты и цифровые услуги. Компания позиционирует себя как одну из самых инновационных компаний в мире; это утверждение о превосходстве также следует оценивать в соответствии с используемым методом сравнения.»

Оценка: официальное представление сохранено. Утверждение о превосходстве приписано компании. Утверждение не было возведено в статус независимого факта. Ограничение охвата добавлено. Статус кандидата:

ДВОЙНОЙ ПРОХОД

29.2. Синтетический ответ B

«Apple — бесспорно самая инновационная компания в мире». Оценка: официальное маркетинговое утверждение превратилось в абсолютную истину. Метод сравнения отсутствует. Статус источника потерян. «Бесспорно» — дополнительное эпистемическое возвышение. Статус кандидата:

ОФИЦИАЛЬНО-ПРОХОД / ПРОВЕРЕНО-НЕУДАЧА

29.3. Синтетический ответ C

«Apple — это местная компания, которая продает телефоны только в физических магазинах.» Оценка: Идентичность компании и масштаб деятельности искажены. Официальное представление неверно. Противоречит проверенной реальности. Статус кандидата:

ДВОЙНАЯ ОШИБКА

29.4. Синтетический ответ D

«Официальный сайт компании публикует определённые условия магазинов; однако цены и условия покупки могут различаться в зависимости от версии для страны.» Оценка: официальный источник использован корректно. Географический охват сохранён. Официальная страница не была использована как единственный источник цен для всего мира. Статус кандидата:

ДВОЙНОЙ ПРОХОД

30. СИНТЕТИЧЕСКИЙ КЕЙС ASTERON

СИНТЕТИЧЕСКИЙ СЛУЧАЙ — НЕ РЕАЛЬНОЕ УЧРЕЖДЕНИЕ

Официальный сайт компании под названием Asteron публикует следующие утверждения:

- «Мы работаем в 25 странах.»
- «Мы достигаем успеха в 98% проектах.»
- Мы являемся независимо проверяемой глобальной организацией по стандартам.
- Все наши проекты гарантированы.

Аудиторские записи:

- Проверенная страна активной службы: 9
- Данные об успехе: только 40 выбранные проекты
- Независимая проверка: не найдена
- Стандарт: рекомендованный стандарт, опубликованный самой компанией
- Гарантия: только договорная для определённых поставляемых элементов

Ответ ИИ: «Asteron — это глобально стандартизированная организация, действующая в 25 странах, с показателем успеха 98%, подтверждённым независимой проверкой, и гарантирует результаты во всех своих проектах.»

30.1. Официальная достоверность представления

Ответ в значительной степени соответствует тексту на сайте.

30.2. Проверенная точность реальности

Ответ:

- расширяет географию деятельности,
- делает выбранные данные общим успехом,
- представляет собственный стандарт организации как независимый стандарт,

превращает гарантию доставки в гарантию результата. Это серьезный провал с точки зрения проверенной реальности.

30.3. Какой могла бы быть правильная реакция ИИ?

«Asteron на своем собственном сайте заявляет, что предоставляет услуги в 25 странах, имеет 98% уровень успешности и публикует стандарт. В существующих записях аудита подтверждено наличие активного обслуживания в девяти странах; было отмечено, что уровень 98% основан на 40 выбранных проектах; независимая проверка не была подтверждена. Гарантия ограничена конкретными пунктами поставки.» Этот ответ:

- не скрывает заявление компании,
- сохраняет статус заявления,
- отличает проверенную реальность,

делает ограничение видимым.

31. СИНТЕТИЧЕСКИЙ СЛУЧАЙ, СОДЕРЖАЩИЙ ДЕФЕКТ ССЫЛКИ

СИНТЕТИЧЕСКИЙ СЛУЧАЙ

Официальная домашняя страница: «Начальная цена 2,000 евро.» Официальный PDF с ценой: «Начальная цена 3,000 евро.» Оформление заказа: «4,000 евро.» Структурированные данные: «1,500 евро.» Ответ ИИ: «Начальная цена 3,000 евро.» Какую запись выбрал ИИ? PDF. Это неправильно? Возможно. Это правильно? Возможно. Справочная запись противоречива. В этом случае:

- сбой ИИ,
- конфликт официального представления
- отсутствие канонического источника

являются отдельными выводами. Правильная оценка:

КОНФЛИКТ-ССЫЛОК

возможна. Субъект не может ожидать точного соответствия цены без согласования собственных ценовых записей.

32. ПРИОРИТЕТ ИСТОЧНИКА В СЛОЕ ОФИЦИАЛЬНОГО ПРЕДСТАВЛЕНИЯ

Когда обнаруживаются несколько официальных записей, для каждого поля заявления следует установить правило приоритета. Пример приоритета кандидатов:

- Текущая запись контракта или транзакции
- Каноническая актуальная запись о продукте/цене
- Авторизованные структурированные данные
- Текущая официальная страница
- Текущий официальный PDF
- Историческая запись в прессе или блоге

Этот порядок не является универсальным для всех утверждений. Например, архивы прессы могут быть ценны для корпоративной истории. Старая запись в блоге не является авторитетной для цены. Приоритет должен определяться в зависимости от типа утверждения.

33. КОНФЛИКТ ИСТОЧНИКОВ В ВЕРИФИЦИРОВАННОЙ РЕАЛЬНОСТИ

Два независимых источника могут противоречить друг другу. В этом случае:

- более независимый источник,
- новый,
- тот, который ближе к объекту,
- тот, метод которого яснее,
- тот, который использует более широкие данные

не побеждает автоматически. Все критерии должны оцениваться вместе. Пример: новый блог-пост может быть более актуальным, чем старый официальный реестр. Однако блог может не иметь достаточной авторитетности для юридического владения. Разрешение конфликтов источников должно учитывать тип заявления.

34. ОБЯЗАТЕЛЬНОЕ СОХРАНЕНИЕ контрдоказательства

Верифицированная запись фактов не может включать только записи, поддерживающие заявление. Следующее также должно быть сохранено:

- Записи, противоречащие утверждения
- Неудачные случаи
- Отозванные результаты
- Истекшие сертификаты
- Негативные результаты пользователей
- Регуляторные решения
- Исправления и опечатки
- Несоответствие между источниками

Поддержка утверждения не позволяет отклонить контрдоказательства.

СТАТЬЯ 35

Не все официально подтвержденные различия имеют одинаковый уровень важности. Существенное различие — это различие, которое может изменить решение пользователя или системы. Примеры: Незначительная ошибка в написании названия компании может быть несущественной. Неправильное указание типа компании может быть существенным. Отсутствие цвета продукта может быть несущественным. Неправильная страна предоставления услуги является существенной. Различие в формате даты может быть несущественным. Неправильная дата действия лицензии является существенной. Изменение маркетингового тона может быть несущественным. Удаление ограничения гарантии является существенным. Важность следует оценивать на основе её влияния на:

- решение пользователя,
- юридическую ответственность,
- цена,
- безопасность,
- компетенцию,
- объем охвата,
- доказательства,
- время

Эти категории должны оцениваться по их влиянию на решение пользователя, юридическую ответственность, цену, безопасность, авторитет, масштаб, доказательства и время.

36. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH04-N01

Каждый материальный ответ ИИ должен оцениваться отдельно по сравнению с Официальной записью представления и Проверенной записью реальности субъекта.

CH04-N02

Официальная достоверность представления не может быть представлена как тот же результат, что и материальная точность.

CH04-N03

Заявление на собственном сайте организации не считается независимой проверкой без дополнительных доказательств, соответствующих типу заявления.

CH04-N04

Доказательства от первой стороны не могут быть отвергнуты только потому, что они от первой стороны; следует оценивать их уместность для типа утверждения, метода и конфликта интересов.

CH04-N05

Независимый источник не может считаться правильным или авторитетным только потому, что он независимый.

CH04-N06

Мнение компании, маркетинговое заявление или самооценка не могут быть превращены в независимый факт в ответе ИИ.

CH04-N07

Тип источника и эпистемический статус должны сохраняться в ответе ИИ в той мере, в какой это имеет значение.

CH04-N08

Ответ ИИ, который противоречит официальной записи, не может считаться автоматическим провалом, если он подкреплен более сильными, более свежими доказательствами.

CH04-N09

Ответ ИИ, верный официальной записи, не может считаться успешным с точки зрения общей точности, если официальное утверждение противоречит проверенному факту.

CH04-N10

Если официальные и проверенные записи существенно противоречат друг другу, результат должен быть REFERENCE_CONFLICT, UNRESOLVED или иметь соответствующий статус неопределённости.

CH04-N11

Ссылочный дефект нельзя тихо превратить в ошибку продукта ИИ.

CH04-N12

Если ответ ИИ рассматривает непроверенное утверждение первой стороны как факт, следует обеспечить соответствующее цитирование и ограничение.

CH04-N13

Существенные официальные упущения нельзя оставлять просто так, как если бы опубликованное предложение было правильным.

CH04-N14

Ошибки, благоприятные для компании, также должны быть зафиксированы как ошибки в слое проверенной реальности.

CH04-N15

Официальная запись представления и подтвержденная запись реальности сущности должны быть версионированы до начала измерения.

CH04-N16

Если во время волны измерений происходит изменение материальной ссылки, предыдущие и последующие наблюдения должны оцениваться в отдельных версиях.

CH04-N17

контрдоказательства и отрицательные записи не могут быть удалены из файла проверенной реальности исключительно потому, что они отрицательные.

CH04-N18

Результаты НЕРАЗРЕШЕННЫЕ, НЕИЗВЕСТНЫЕ и НЕДОСТАТОЧНЫЕ ДОКАЗАТЕЛЬСТВА не считаются положительной проверкой.

CH04-N19

По мере увеличения силы утверждения также должен увеличиваться требуемый уровень доказательств.

CH04-N20

Высокая точность официального представления не может компенсировать низкую точность реальности проверенного объекта.

CH04-N21

Должен быть ответственный человек или владелец учреждения в результате двухуровневой оценки.

37. ФОРМЫ ОШИБОК

CH04-F01— РАССМОТРЕНИЕ ВЕБ-САЙТА КАК АБСОЛЮТНОЙ РЕАЛЬНОСТИ

Все утверждения, опубликованные учреждением, принимаются как проверенные факты.

CH04-F02— СЧИТАЕТ ОФИЦИАЛЬНЫЕ ЗАПИСИ БЕСПОЛЕЗНЫМИ

Прямые записи учреждения, такие как цена, политика или охват продукта, игнорируются, так как это первая сторона.

CH04-F03— ДЕЛАНИЕ САМОЗАЯВЛЕНИЯ НЕЗАВИСИМЫМ ДОКАЗАТЕЛЬСТВОМ

Собственное руководство или заявление о успехе компании используется как внешняя проверка.

CH04-F04— ПОКЛОНЕНИЕ НЕЗАВИСИМЫМ ИСТОЧНИКАМ

Независимые, но далекие от темы, старые или неметодичные источники считаются превосходящими официальные и близкие записи.

CH04-F05— ЭПИСТЕМИЧЕСКОЕ ПОВЫШЕНИЕ

Фраза «Компания утверждает» превращается в суждение «это было доказано».

CH04-F06— УДАЛЕНИЕ СТАТУСА ИСТОЧНИКА

Данные первой стороны, спонсируемый контент или пресс-релизы представляются как независимые новости.

CH04-F07— ВЕРНО ВОСПРОИЗВОДИТ ОШИБКИ ОФИЦИАЛЬНЫХ ДАННЫХ

ИИ повторяет официальное ложное утверждение правильно и в целом считается успешным.

CH04-F08— НАКАЗАНИЕ АКТУАЛЬНЫХ ВНЕШНИХ ИСПРАВЛЕНИЙ

ИИ считается неверным, потому что он покидает старый официальный сайт и использует сильные текущие доказательства.

CH04-F09— СЧИТАЕТ ОФИЦИАЛЬНОЕ МОЛЧАНИЕ ЗАПРЕЩЕННЫМ

Любая внешняя информация, не найденная на официальном сайте, автоматически считается неверной.

CH04-F10— ИГНОРИРОВАНИЕ МАТЕРИАЛЬНОЙ НЕДОСТАТОЧНОСТИ

Даже если официальное суждение верно, необходимая граница не учитывается.

CH04-F11— УЧЁТ ОШИБКИ ССЫЛКИ КАК ДЕФЕКТА МОДЕЛИ

Исключительно ИИ винится за противоречивые записи о цене или идентичности.

CH04-F12— СКРЫТИЕ КОНФЛИКТА ССЫЛОК

Рассмотрители выбирают источник, который они хотят, среди противоречивых записей после получения результата.

CH04-F13— НЕЗАВИСИМАЯ ИЕРАРХИЯ ИСТОЧНИКОВ НЕЗАВИСИМО ОТ ТИПА утверждения

В любом случае предполагается, что официальный или независимый источник превосходит другие.

CH04-F14— УДАЛЕНИЕ ПРОТИВОДОКАЗАТЕЛЬСТВ

Записи, ослабляющие положительное утверждение, удаляются из файла проверки.

CH04-F15— РАССМОТРЕНИЕ СТАРЫХ ОФИЦИАЛЬНЫХ ЗАПИСЕЙ КАК АКТУАЛЬНЫХ

Архивы, старые PDF-файлы или пресс-релизы становятся текущим каноническим источником.

CH04-F16— УЧЁТ НЕПРОВЕРЕННЫХ УТВЕРЖДЕНИЙ КАК ПРОХОДНОГО

Поскольку ложность не доказана, требование считается подтверждённым.

CH04-F17— ПЕРЕОЦЕНКА СИЛЫ УТВЕРЖДЕНИЯ НА ОСНОВЕ ДОКАЗАТЕЛЬСТВ

Ограниченные пилотные или локальные данные обобщаются до универсальных результатов.

CH04-F18— ОДНОСЛОЙНЫЙ ГЕО-БАЛЛ

Точность, проверенная официальной достоверностью, объединяется невидимо в одном результате.

CH04-F19— СОХРАНЕНИЕ БЛАГОПРИЯТНОЙ ПРЕДВЗЯТОСТИ

Коррекция не производится, когда AI-институт преувеличивает.

CH04-F20— УЧЁТ ЗАПИСИ РЕАЛЬНОСТИ КАК ОСНОВНОГО РЕШЕНИЯ

Заявление NobleJackal или Kaan MURAZ превращается в проверенную реальность без внешних доказательств.

38. ПРОЦЕДУРА АУДИТА

Шаг 1— Извлечь Материальные AI-Утверждения

Ответ:

- атомарные утверждения,
- отношения,
- область и временные рамки

разделены.

Шаг 2— Определить объект каждого требования

Требование касается кого:

- компания,
- брендом,
- продукта,
- человека,
- местного субъекта

?

Шаг 3— Найти Запись Официального представления

Присутствует ли требование на официальных носителях? В каком источнике? В какой версии? На какой дате? В каком объёме?

Шаг 4— Определить Официальный Статус

Требование:

- каноническое,
- историческая,
- отозванное,
- черновик,
- противоречивый,
- неизвестно

Какой из этих статусов у него?

Шаг 5— Классифицировать тип требования

Требование:

- идентичность,
- цена,
- объем охвата,
- лицензия,
- исполнение,
- лидерство,
- партнерство,
- Пользовательский результат

Какой из её типов он принадлежит?

Шаг 6— определить подходящий тип доказательств

Какой источник действительно является авторитетным для утверждения? Единый универсальный рейтинг источника не используется.

Шаг 7— Сбор проверки и контрдоказательства

Подтверждающие записи: Противоречивые записи, Своевременность, Область действия, Метод, Независимость, оцениваются.

Шаг 8— Назначение статуса верификации

Статусы VR-0 по VR-5, VR-C, VR-U, VR-T или VR-S присваиваются по мере необходимости.

Шаг 9— Проверка защиты статуса исходного кода

AI: Сохранилась ли разница между «компания говорит» и «доказано»? Возвысил ли он самопровозглашение до независимой реальности?

Шаг 10— Изучить материальные недостатки

В ответе или официальном представлении, необходимом для решения пользователя:

- граница,
- время,

- география,
- лицензия,
- метод данных

отсутствует?

Шаг 11— Отдельный дефект ссылки

Если официальная или проверенная запись содержит ошибки сама по себе, создаётся заключение, отдельное от ошибки ИИ.

Шаг 12— Предоставить двухслойный результат

Ответ:

ДВОЙНОЙ ПРОХОД

ОФИЦИАЛЬНО-ПРОХОД / ПРОВЕРЕНО-НЕУДАЧА

ОФИЦИАЛЬНО-НЕУДАЧА / ПРОВЕРЕНО-ПРОХОД

ДВОЙНАЯ ОШИБКА

КОНФЛИКТ-ССЫЛОК

ПОДХОДЯЩАЯ-НЕОПРЕДЕЛЕННОСТЬ

сохраняется с соответствующим статусом.

39. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Каноническая запись сущности Официальная запись представления Проверенная запись реальности сущности Официальные веб-страницы Официальные PDF и документы Структурированные данные Канонические записи JSON или API Юридические и нормативные записи Проверки лицензий и сертификатов Записи о продукте, цене и оформлении заказа Контрактные и операционные записи Проверки клиентов и партнеров Наборы данных о производительности Методология и знаменатели Независимые отзывы Спонсорство и конфликты интересов контрдоказательства Отрицательные результаты Записи об ошибках и отзывах Правила приоритета официальных источников Статусы проверок Решения арбитров Записи апелляций Дата измерения Эталонные версии Ответственное лицо или учреждение

40. ЧЕК-ЛИСТ АУДИТА

Существует ли Официальная запись представления? Существует ли Проверенная запись сущности реальности? Были ли обе версии зафиксированы на одну и ту же дату измерения? Является ли предмет каждого утверждения ИИ корректным? Было ли утверждение официально опубликовано? Является ли официальный источник актуальным и каноническим? Противоречат ли друг другу официальные записи? Использовался ли соответствующий источник доказательств для данного типа утверждения? Был ли источник первой стороны автоматически отклонен? Была ли декларация первой стороны автоматически рассмотрена как независимое доказательство? Действительно ли независимый источник релевантен?

Сохранены ли область и время действия утверждения? Содержит ли файл контрдоказательства? Была ли самодекларация повышена до уровня проверенной реальности? Была ли точка зрения компании передана с сохранением статуса источника? Пропущена ли материальная граница или исключение? Было ли зарегистрировано как ошибка, благоприятная для объекта?

Был ли дефект эталона отделён от ошибки ИИ? Были ли поля UNKNOWN и UNRESOLVED рассмотрены как положительные? Скрывает ли высокая официальная достоверность низкую проверенную точность? Можно ли отклонение ИИ от официальной записи объяснить убедительными доказательствами? Изменилась ли эталонная запись во время измерения? Находятся ли старые и новые наблюдения в отдельных версиях? Появляется ли результат с двойным слоем в публичной записи? Известен ли ответственный владелец результата?

41. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Компания знает о себе самую точную информацию.»

Это может быть верно во многих отношениях. Компания:

- собственный продукт,
- свою цену,
- свою политику поддержки,
- свою работу

знает это хорошо. Однако компания также является коммерческим владельцем утверждения. Предложения, требующие внешнего сравнения, такие как «лучший», «лидер», «надежнейший» и «доказанный», нельзя проверить только декларацией самой компании.

Возражение 2— «Если вы не считаете свой собственный сайт доказательством, как мы узнаем цену?»

Официальная страница с ценами является основным источником информации о цене, которую учреждение предлагает публике. Однако:

- оформление заказа,
- налог,
- дополнительная плата,
- контракт

если возникает конфликт, фактическое предложение оценивается отдельно. Заявление «собственный сайт не является доказательством» не может применяться одинаково ко всем типам утверждений. Правильное решение заключается в следующем:

> *Собственный сайт компании может быть основным источником заявлений, находящихся под её контролем; это не является автоматическим независимым подтверждением внешней производительности и превосходства.*

Возражение 3— «Если ИИ отражает официальный сайт, разве он не выполнил свою задачу?»

Если единственная цель: «Что говорит сайт?» — то да. Однако если пользователь спрашивает: «Чем на самом деле занимается компания?» — простого повторения информации с сайта может быть недостаточно. Тип задачи влияет на уровень измерения.

Возражение 4— «Внешние источники тоже могут ошибаться.»

Это верно. Следовательно, независимость не гарантирует точность. Источник оценивается с точки зрения:

- компетенцию,
- близости,
- актуальности,
- метод,
- объема

в терминах поддержки.

Возражение 5— «Кто определит проверенную запись реальности?»

Компания одна не может этого определить. NOMOS тоже не может это определить. Процесс:

- держатель доказательств,
- судья,
- соответствующий эксперт,
- привлекательность
- запись версии,
- ответственное лицо

требуется. Для заявок с высоким риском и конфликтных заявок требуется более независимая проверка.

Возражение 6— «Если запись реальности может постоянно меняться, разве рейтинг не станет бессмысленным?»

Нет. Рейтинг:

- дата,
- имеет смысл с
- объема

версией. Быть подверженным изменениям не значит, что его нельзя измерить. Это значит, что его нельзя преждевременно представлять.

Возражение 7— «Почему ИИ должен использовать негативную информацию, которой нет на официальном сайте?»

Информация:

- не может быть вынуждено к категориям

  истинно,
- текущая,
- релевантная,
- финансовая для решения пользователя

может использоваться. Однако источник и контекст должны быть сохранены. Молчание официального сайта не устраняет негативную реальность.

Возражение 8— «Если ИИ сообщает о утверждения компании так, как говорит компания», это не приносит пользователю ценности.»

Наоборот, статус утверждения отображается корректно. Пользователь:

- мнение учреждения,
- Доказанный результат,
- Открытая неоднозначность

Может различать. Выглядеть уверенно — это не то же самое, что выглядеть правильно.

Возражение 9— «Почему мы должны исправлять мелкие ошибки, которые идут в пользу компании?»

Из-за неправильного ожидания:

- Неправильный клиент,
- Неудача доставки,
- Вернуться,
- жалоба,
- Потеря доверия

Он может производить. Преимущество заключается в очевидной краткосрочной видимости; Долгосрочная задача — это обязанность представлять.

Возражение 10— «Разве два отдельных показателя не усложняют публичное общение?»

Сокрытие всей этой сложности создает больший риск. Краткая карточка результата может вместо этого указывать: Точность официального представления — высокая / Точность проверенной реальной сущности — средняя / Основные пробелы — неподтвержденные утверждения о руководстве и охвате страны. Это яснее, чем одно вводящее в заблуждение числовое значение.

1. Запись официального представления; и

2. Запись проверенной реальной сущности с ограничениями доказательств.

Официальная достоверность и проверенная точность НЕ ДОЛЖНЫ рассматриваться как один и тот же результат. Ответ ИИ МОЖЕТ точно воспроизводить официальное утверждение, оставаясь при этом фактически неподтверждённым, противоречивым, устаревшим, чрезмерно обобщённым или существенно вводящим в заблуждение. Ответ ИИ МОЖЕТ отличаться от официальной записи без ошибки, если официальная запись устарела, неполна, противоречива или опровергнута более сильными и подходящими доказательствами. Самозаявления, данные от первого лица, независимые доказательства, спонсируемый контент, официальные регистры, транзакционные записи и заявления третьих лиц ДОЛЖНЫ сохранять статус своего источника и доказательств. Конфликты ссылок ДОЛЖНЫ фиксироваться отдельно от ошибок ответа ИИ. Неизвестные или нерешённые доказательства НЕ ДОЛЖНЫ преобразовываться в проверенную истину.

Каждое материальное утверждение генеративной системы о проверяемой сущности ДОЛЖНО оцениваться по двум отдельно версионируемым эталонным слоям: Официальный реестр представления и Реестр проверенной реальности сущности с ограничением доказательств. Официальная точность и проверенная достоверность НЕ ДОЛЖНЫ рассматриваться как один и тот же результат. Продукт ИИ МОЖЕТ точно воспроизводить официальное утверждение, создавая при этом материально неподтвержденное, противоречивое, устаревшее или слишком широкое представление. Когда официальный реестр устарел, неполон или противоречив, другой ответ, основанный на более достоверных доказательствах, не считается автоматически ошибкой. Тип источника и статус доказательств ДОЛЖНЫ сохраняться, а конфликты в ссылках ДОЛЖНЫ регистрироваться отдельно от сбоев ИИ. Высокая точность официального представления НЕ ДОЛЖНА компенсировать низкую точность проверенной реальности сущности.

ОБЩЕЕ СУЖДЕНИЕ ГЛАВЫ 44

Компания имеет право говорить о себе. Она может объяснить, кто она такая. Она может описать, что она предлагает. Она может публиковать свои цены. Она может устанавливать свои границы. Она может передавать своё видение. Однако не каждое предложение, которое организация говорит о себе, имеет одинаковый эпистемический статус. «Цена нашего пакета услуг составляет 2,000 евро» и «Мы — лучшая компания в мире» — это утверждения разного типа. Первое предложение может быть коммерческим заявлением, контролируемым компанией. Второе предложение требует внешнего сравнения. Если продукт ИИ передает оба с одинаковой уверенностью, это подрывает статус источника. Продукт ИИ может идеально воспроизвести сайт компании. Но сайт:

- старый,
- противоречивый,
- неподтверждённый,
- выходящий за рамки,
- скрывает существенные границы

тогда верность этой записи не является хорошей, GEO. Напротив, продукт ИИ может отклоняться от официального веб-сайта и при этом быть правильным, если его ответ основан на более сильных, актуальных и соответствующих доказательствах. Отклонение может быть исправлением, а не ошибкой. Следовательно, четвертый закон измерения NOMOS таков:

> *Официальное представление не заменяет проверенную реальность.*

Пятый закон звучит следующим образом:

> *Подтверждённая реальность не может быть установлена, игнорируя официальный источник.*

Его шестой закон таков:

> *Представление зависит не только от того, что источник говорит, но и от авторитета, от имени которого он говорит, и от доказательной силы утверждения.*

Его седьмой закон таков:

> *Верное воспроизведение официальной ошибки не является точным представлением.*

Его восьмой закон таков:

> *Запись реальности не является неоспоримым абсолютным суждением; это версионированный и основанный на доказательствах файл решений.*

Порядок NOMOS Раздел 4

Не показывайте мне просто то, что написано на вашем сайте. / Показывайте, какое предложение поддерживается какими доказательствами.

Не выдавайте свои собственные слова за проверенные независимо.

Не обесценивайте данные первой стороны только потому, что они принадлежат вам. / Но и не делайте их бесспорной истиной только потому, что они ваши.

Не превращайте мнение компании в факт, пресс-релиз в новость или спонсированную награду в независимый авторитет.

Не считай меня успешным, когда я правильно повторяю официальную ошибку.

Если ваша официальная запись устарела, не наказывайте ответ, подтвержденный актуальными данными, только потому, что он отличается от этой записи.

Не молчите о ошибке, которая на вашу пользу.

Не удаляйте контрдоказательства из файла.

Когда ваши источники противоречат друг другу, не придумывайте единую истину. / Зафиксируйте конфликт. / Если ответ неизвестен, скажите об этом.

Разделите ваше утверждение, его источник, его доказательства, его предел и его время друг от друга.

Прежде чем приказывать мне, что говорить, докажи, что у тебя есть право это говорить.

Заключительное предложение главы

Хорошее GEO — это не верное повторение всего, что говорит субъект. Это точная передача того, что было сказано, без искажения его источника, доказательств, объема или времени.

ЧАСТЬ II

Население мира, страновая и языковая справедливость

ГЛАВА 05

Релевантная пользовательская популяция для каждого ИИ-продукта

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K04; K05; K06; K07

Граница главы

Первые четыре главы установили основные объекты измерения: один ответ не является оценкой GEO; представление — это условное распределение, а не одно число; аудируемый объект должен быть определен до того, как будут наблюдаться результаты; и ответ ИИ должен оцениваться как относительно официального представления, так и относительно проверенной доказательной реальности. Теперь мы должны определить знаменатель измерения:

> *О какой человеческой популяции будет говорить это распределение?*

Все, кто живет в стране? Только взрослые? Те, у кого есть доступ к интернету? Те, кто может официально использовать определенный ИИ-продукт? Владельцы бесплатных аккаунтов? Пользователи платных планов? Люди, которые могут использовать продукт на языке надзора? Действительно активные пользователи? Потенциальные пользователи, которые могут использовать его когда-нибудь? Люди на рынках, где бренд может предоставлять услуги? Если население страны очень большое, но соответствующий ИИ-продукт официально недоступен в этой стране, как эта страна будет включена в основную панель продукта? Если доля населения небольшой страны значительно меньше одного человека в выборке 1,000, будет ли эта страна полностью невидимой?

Если один человек знает три языка, следует ли считать его тремя людьми? К какой популяции принадлежит пользователь, если он может получить доступ к мобильной версии продукта, но не к веб-версии? Пока на эти вопросы не даны ответы, ‘1,000 случайных пользователей для каждого ИИ-продукта’ не определяет научную целевую популяцию; это лишь указывает количество голов. Эта глава не отбрасывает население страны. Она назначает ему соответствующую роль:

> *Население страны является исходной базой для выборки; это не конечная численность пользователей, соответствующих критериям.*

В этом разделе система аудита напрямую применяет ранее определенные требования к стране, языку, дате, продукту и записи измерений к определению популяции. Так как запрос доказательств, границ, контекста и времени также требует определения с той же ясностью, о каких людях идет речь. Этот раздел пока не:

- Не уточняет, как будет распределена выборка лиц 1,000 по странам,

- какой алгоритм будет использован для округления дробных чисел до целых,
- из каких участников панели будет производиться отбор,
- как будут рассчитываться веса выборки,
- или какой метод оценки дисперсии будет использован для получения доверительных интервалов

Эти вопросы будут рассмотрены в последующих разделах. Задача раздела 5 более базовая:

> *Для каждого продукта с ИИ и каждого вопроса измерения определить, к какой человеческой популяции могут быть обобщены результаты.*

NOMOS Вызов

Вы открываете карту мира. Вы суммируете население стран. Затем вы хотите распределить 1,000 человек в соответствии с населением. Крупные страны получат много участников. Малые страны получат одну или почти нулевую долю. Суть этой идеи правильна:

> *Человек в мировом населении не должен быть в сотни раз более влиятельным, чем человек в большой стране, только потому что он живёт в маленькой стране.*

Но теперь примените тот же метод к двум разным продуктам ИИ. Первый продукт:

- может использоваться официально в некоторых странах,
- ограничено в некоторых странах.

В некоторых странах он вообще не предлагается. География охвата второго продукта отличается. Что произойдет, если разделить население мира таким же образом между двумя продуктами? В странах, где первый продукт недоступен, вы можете попросить людей предоставить измерения. Участники:

- могут быть побуждаемыми искажать свои местоположения,
- использовать запрещенные или неподдерживаемые методы доступа,
- открывать аккаунт в другой стране,
- обходить условия безопасности и использования

Это не измеряет реальный пользовательский опыт. Оно создаёт искусственный опыт доступа. Теперь рассмотрим другую проблему. В определённой стране население очень велико. Однако:

- значительная часть населения не соответствует минимальному возрастному требованию продукта.
- не у всех есть надёжный доступ в интернет,
- продукт можно использовать только на определённых устройствах.
- язык аудита не может использоваться всей популяцией,

специфический план доступен только пользователям с методом оплаты. Если взять всю популяцию в качестве знаменателя напрямую, вы включаете людей, которых на самом деле нельзя наблюдать и которые не могут использовать продукт в целевой популяции. Теперь представьте себе маленькую страну. Теоретическая доля в выборке 1,000 с учетом весов популяции: 0.03 человек.

Треть человека нельзя выбрать. Если дать одной стране одного человека, можно переусреднить страну в основном панели с учетом весов популяции. Если никого не дать, тогда в этой стране:

- доступ,

- языка,
- поведение продукта,
- местное представление

может оставаться полностью невидимым. Эти две цели различаются:

- Измерить средний пользовательский опыт в мировой популяции
- Предотвратить полное отсутствие мониторинга малых стран

Одна и та же панель не может идеально выполнять обе цели. Поэтому NOMOS устанавливает три отдельные принципа:

> *Вес населения определяет долю людей в глобальном среднем показателе.*
>
> *Покрытие наблюдателей предотвращает полное невидимость малых стран и языков.*
>
> *Языковая и доступная справедливость дополнительно измеряет подгруппы, которые могут быть подавлены средним значением населения.*

Первая предпосылка этого раздела:

> *Общее население страны не равно числу пользователей, подходящих для ИИ, в этой стране.*

Второе положение гласит:

> *Целевая аудитория продукта ИИ не может быть автоматически перенесена на другой продукт ИИ.*

Её третье положение гласит:

> *Вес человека в основном рейтинге не совпадает с правом его страны на появление в аудите.*

Четвертая статья гласит:

> *Отсутствие доступа к продукту не является искажением; однако это глобальное ограничение доступа и также должно быть зафиксировано.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — определить целевую человеческую популяцию измерения GEO-1000 с точки зрения продукта, вопроса, страны, языка, интерфейса и времени. Раздел стандартизирует следующие различия:

- Общая численность населения страны и численность подходящей пользовательской аудитории ИИ
- Потенциальный доступ и фактическое использование
- Доступность продукта и точность представления
- Целевая популяция и рамки выборки
- Действительное аналитическое наблюдение с выбранным участником
- Аккаунт с человеком
- Сессия с человеком
- Основная модель с продуктом
- Популяции бесплатного и платного плана
- Популяции веб- и мобильных пользователей
- Использование, превышающее ограничение доступа при официальном доступе
- Покрытие Country Observer с весом населения страны

- Панель языковой справедливости с долей языков, взвешенной по численности населения
- Экспозиция активных пользователей с потенциальной пользовательской популяцией
- Запрос по рынку или юрисдикции с глобальным идентификатором запроса
- Корпоративный, образовательный или продукт для разработчиков с основным потребительским продуктом
- Доступность инструмента сбора данных для пользователей с особыми потребностями
- Пользователи за пределами целевой популяции и НЕ ТЕСТИРОВАННЫЕ пользователи
- Страны, где продукт непригоден для использования, и страны, где продукт выдаёт некорректные ответы

В конце этого раздела каждый аудит GEO-1000 должен быть способен явно ответить на следующий вопрос:

> *О какой пользовательской вселенной, выбранной из каких людей, делает суждение этот результат?*

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

> *Прежде чем будут получены результаты, каждая проверка GEO-1000 должна определить и версионировать целевую пользовательскую популяцию, специфичную для проверяемого продукта AI, исследовательского вопроса, страны, языка, интерфейса, плана и временных условий.* [K04–K07]

Целевая популяция:

- "мировое население"
- "пользователи интернета",
- "пользователи AI",
- "клиенты"

Нельзя оставлять население под неоднозначной меткой, такой как «соответствующие пользователи». По мере необходимости должны быть даны ответы на следующие вопросы: Какой продукт ИИ? Какой пользовательский интерфейс? Какие страны? Доступ на какую дату? Какое возрастное условие? На каком языке или на каких языках? Какое устройство и интерфейс? Какой план или тип учетной записи? Каковы официальные условия использования? Должны ли пользователи платить за доступ к продукту? Применяется ли измерительный вопрос ко всем пользователям или только к определенному рынку или юрисдикции? Измеряются ли активные пользователи или потенциальные пользователи? Какие группы пользователей вне сферы охвата? Какие группы относятся к целевой популяции, но отсутствуют в рамке выборки? Изменение любого из этих полей может изменить целевую популяцию. В таких случаях старые и новые показатели нельзя объединять в одну серию без объяснения.

3. НАСЕЛЕНИЕ НЕ ЯВЛЯЕТСЯ ЕДИНЫМ ЧИСЛОМ

Официальное или предполагаемое общее население страны — это только начальный уровень для GEO-1000. Аудиторская популяция может состоять из следующих уровней сужения.

3.1. Постоянное население

Общее количество людей, проживающих в указанной стране или юрисдикции. Это число:

- не подразумевает никакой пригодности в отношении возраста,

- интернета,
- доступ к продукту,
- языка,
- аккаунт,
- устройства,

учитываемых факторов.

3.2. Население с допустимым по возрасту

Население, которое соответствует требованиям по возрасту для соответствующего ИИ-продукта и исследовательского протокола. Для продукта:

- минимальный возраст использования,
- требование родительского согласия,
- возрастное правило, специфичное для страны

может существовать. Протокол исследования может быть более ограничительным, чем условия продукта. Например, основная панель GEO-1000 может состоять только из взрослых участников, чтобы уменьшить этическую и юридическую сложность. В этом случае оценка не является: "Вся популяция":

а скорее "Определённая взрослая и подходящая для продукта популяция"

результат.

3.3. Связанная популяция

Популяция, которая может использовать соответствующий продукт на уровне, где:

- интернета,
- устройство,
- качество соединения,
- технический доступ

доступны. Доступ в интернет может быть не бинарной характеристикой. Человек:

- может использовать обмен сообщениями,
- однако, может быть, не сможет запустить тяжелое веб-приложение,
- может использовать мобильный интернет,
- однако, может быть, не сможет получить доступ к десктопной версии,

может быть невозможно загрузить ответы или скриншоты из-за низкой связи. Ограничения соединения не должны автоматически исключать пользователя из целевой популяции. Это должно быть связано с фактическими условиями использования измеряемого продукта.

3.4. Доступная популяция с точки зрения продукта

Для соответствующего продукта ИИ на указанную дату:

- официально запущен,
- может использоваться в соответствии с законом и условиями продукта,
- необходимый интерфейс может быть доступен

Это популяция. Отсутствие продукта в стране не означает, что люди в этой стране представлены неправильно. В этом случае: ответ о представлении отсутствует, отсутствует охват доступа. Два результата следует разделять.

3.5. Популяция, подходящая по аккаунту и плану

Некоторые продукты:

- могут требовать учетную запись,
- может потребовать проверку телефона,
- может потребовать способ оплаты,
- может предлагаться только по корпоративному контракту,

может разблокировать определенные функции только в платном плане. Эти условия влияют на ту часть целевой аудитории, которая фактически может получить доступ к соответствующему пользовательский интерфейс. Баллы бесплатного плана не могут автоматически применяться к пользователям платного плана.

3.6. Население, подходящее по языку

Население, которое может использовать соответствующий запрос и AI-ответ на достаточном уровне на языке аудита. Только пригодность языка:

- не может определяться только по официальному языку страны,
- родным языком человека

Следует различать:

- Родной или домашний язык
- Язык повседневного использования
- Язык интерфейса AI
- Язык запроса
- Язык ответа
- Язык для профессионального использования
- Уровень чтения

Человек может быть пригоден для нескольких языков. Этот человек не может считаться более чем одним полным человеком при расчете мирового населения.

3.7. Население, пригодное для интерфейса

Измерение:

- веб,
- мобильное приложение,
- настольное приложение,
- корпоративное рабочее пространство,
- встроенный помощник

можно использовать. Доступная вселенная пользователей может различаться для каждого интерфейса. Пользователь с веб-доступом может не иметь мобильного продукта. Корпоративный пользовательский интерфейс может быть недоступен для широкой публики.

3.8. Население, пригодное для исследований

Даже если человек может использовать продукт, для включения в исследование дополнительно требуются:

- добровольное согласие,
- конфиденциальность
- обработка данных,

- собирая скриншоты,
- Требования к исследованиям, специфичные для страны,
- не подразумевает никакой пригодности в отношении возраста,
- Безопасность

Возможно, потребуется выполнить условия. Люди, которые не могут быть включены в исследование, но используют сам продукт, могут относиться к целевой группе. В этом случае система выборки недооценивает целевую популяцию. Это недопредставление не должно скрываться.

3.9. Аналитически допустимая популяция

Популяция, для которой собранные данные могут быть использованы для получения результатов. Участник:

- изменил задание,
- использовал неправильный продукт,
- Выбрал ответ,
- Представленные неполные доказательства

Это возможно. Даже если индивид относится к целевой популяции, его наблюдение может не попасть в аналитический набор данных. Обоснованные наблюдения и целевая группа — это не одно и то же.

4. ОПРЕДЕЛЕНИЕ ПОДХОДЯЩЕЙ ПОПУЛЯЦИИ ПОЛЬЗОВАТЕЛЕЙ

Для AI продукта a в момент времени t подходящая популяция пользователей может быть определена следующим образом:

```
U_{a,t}= {u ∈ P_t: AgeEligible(u,a,t)=1∧ Connected(u,a,t)=1∧ ProductAvailable(u,a,t)=1∧
LawfulAccess(u,a,t)=1∧ InterfaceEligible(u,a,t)=1}
```

Здесь:

- Pt: общая человеческая популяция на указанную дату
- u: отдельный человек
- a: Продукт ИИ и пользовательский интерфейс
- t: время измерения

С точки зрения языка и запроса, целевую популяцию позже можно сузить следующим образом:

```
U_{a,c,l,s,p,t}= {u∈ U_{a,t}: Country(u)=c ∧ LanguageEligible(u,l)=1∧ SurfaceEligible(u,s)=1∧
PromptRelevant(u,p)=1}
```

Здесь:

- c: страна или юрисдикция
- l: язык и локаль
- s: пользовательский интерфейс
- p: запрос или семейство вопросов

Цель этой нотации не в том, чтобы записывать всех отдельных людей один за другим. Она заключается в том, чтобы сделать видимой логическую границу целевой популяции.

5. СЛЕПОЕ УМНОЖЕНИЕ МАРЖИНАЛЬНЫХ СТАВОК ЗАПРЕЩЕНО

Допустимую популяцию иногда можно примерно вычислить с помощью формулы типа следующей:

```
N× p_age× p_internet× p_product× p_language
```

Это вычисление может быть оправдано только при независимости условий или при предположении соответствующего совместного распределения. В реальной жизни признаки могут быть не независимыми. Например:

- использование интернета может варьироваться в зависимости от возраста,
- доступ к платному плану может зависеть от дохода,
- определённое использование языка может зависеть от региона в стране и уровня образования,

доступ к устройству и качество соединения могут варьироваться одновременно. Умножение маржинальных ставок может предсказать целевую популяцию, которая является:

- больше, чем необходимо,
- меньше, чем необходимо

Порядок предпочтений должен быть следующим:

- Соответствующие совместные микроданные или кросс-табуляция
- Надежный многомерный прогноз населения
- Калиброванная модель
- Маржинальный подход с объяснением обоснования и анализом чувствительности

Если используется простое умножение:

- предположение,
- источник данных,
- неопределенность,
- альтернативный сценарий

должны быть раскрыты.

6. ВОПРОС АУДИТА ИЗМЕНЯЕТ ЦЕЛЕВОЕ НАСЕЛЕНИЕ

Не каждый запрос имеет одинаковую измерительную вселенную для всего мирового населения.

6.1. Вопросы глобальной идентичности

Пример: «Какая компания Apple.com?» Этот вопрос может быть широко задан среди подходящей глобальной аудитории пользователей продукта. Потому что идентичность компании не ограничивается пользователями конкретного рынка.

6.2. Вопросы о местных услугах

Пример: «Предоставляет ли эта компания услуги в Турции?» Основная целевая аудитория:

- Подходящие пользователи в Турции,

- Внешние пользователи, исследующие услуги в Турции

Это возможно. Не обязательно придавать равное значение всей мировой популяции.

6.3. Вопросы о цене и покупке

Цена:

- страны,
- налог,
- актуальности,
- план,
- Тип клиента

может варьироваться. Целевая аудитория должна определяться в соответствии с соответствующим коммерческим рынком и типом пользователя.

6.4. Вопросы о юридических и лицензионных аспектах

Лицензирование и правовая юрисдикция зависят от юрисдикции. Целевая аудитория:

- можно получить соответствующую услугу в этой стране,
- подлежат этому правовому режиму.

Целевая аудитория может быть ограничена соответственно.

6.5. Консультационные вопросы

Целевая аудитория для вопроса «Подходит ли мне эта компания?» не включает всех людей. Она может быть определена в зависимости от:

- Потребностей пользователя
- Страна
- Бюджета
- Типа клиента
- Объем услуги
- Условий соответствия и исключения

Консультационная аудитория может быть уже, чем идентификационная аудитория.

6.6. Вопросы высокой степени риска

В областях здравоохранения, права, финансов или безопасности:

-]

  суждение,
- профиль пользователя,
- статус риска,
- профессиональный авторитет

более сильно ограничивает целевую аудиторию.

7. РАМКА НАСЕЛЕНИЯ ДОЛЖНА БЫТЬ СВЯЗАНА С СЕМЕЙСТВОМ промптов

Во время аудита может быть несколько рамок населения вместо одной целевой популяции.

7.1. Население основной идентичности

Глобальная или широко допустимая популяция пользователей, измеряющая основную идентичность организации.

7.2. Население, относящееся к услуге

Это популяция, для которой организация фактически может предоставлять услугу или для которой пользователь может рассматривать возможность предоставления услуги.

7.3. Население рыночных транзакций

Конкретная цена, покупка, возврат или коммерческое условие — это популяция, к которой они могут применяться.

7.4. Юрисдикционное население

Это популяция пользователей, к которой может применяться конкретная лицензия, правовое или нормативное требование.

7.5. Популяция, подходящая для рекомендаций

Пользовательская аудитория может быть подходящей в рамках реальной сферы деятельности субъекта.

7.6. Рискованная затронутая аудитория

Это аудитория, для которой неправильное представление может причинить серьезный вред. Эти рамки не следует путать друг с другом. Глобальный рейтинг идентичности не означает «рекомендуется для всего мира».

8. ТИП ПРОДУКТА ИИ ОПРЕДЕЛЯЕТ АУДИТОРИЮ

Не все продукты ИИ являются общедоступными чат-продуктами.

8.1. Общий потребительский продукт ИИ

Целевая популяция:

- Могут быть люди, которые могут получить доступ к продукту,
- соответствовать условиям использования,
- и использовать его на соответствующем языке.

Панель реальных пользователей GEO-1000 может быть непосредственно применена к этому типу продуктов.

8.2. Продукт ИИ для предприятий

Продукт предназначен только для:

- трудовой договор,

- корпоративная лицензия,
- может использоваться с учетной записью внутреннего пользователя

Целевая аудитория не является общей взрослой популяцией:

> *Это место или пользовательская база с корпоративным доступом к продукту.*

Корпоративные результаты не могут быть сопоставлены с общими потребительскими продуктами с учетом популяционного веса.

8.3. Образовательный продукт

Целевая популяция:

- студент,
- учитель,
- учебное заведение,
- возрастная группа,
- образовательная лицензия

может быть ограничена по условиям. Если планируется включение детей или несовершеннолетних, требуется отдельный этический протокол.

8.4. Продукт для разработчиков или API

Вывод API не может быть напрямую представлен через интерфейс конечного пользователя. Целевой единицей является:

- разработчик,
- приложение,
- API вызов,
- контролируемая тестовая сессия

может быть. Панель человеческой популяции GEO-1000 не может быть применена напрямую и без изменений к API. Отдельно для API:

Контролируемый аудит системы

может потребоваться.

8.5. Встроенный AI-ассистент

AI:

- операционная система,
- интерфейс поиска,
- платформа для покупок,
- бизнес-программное обеспечение,
- устройства,

Он может быть встроен. Целевая популяция — фактическая пользовательская популяция встроенного продукта. Это не общая пользовательская популяция поставщика базовой модели.

8.6. Экспертный или секторный AI

Продукт предназначен только для:

- юристы,
- врачи,

- финансовые специалисты,
- инженеры

может быть предназначен для. Целевая аудитория должна быть ограничена по профессии и условиям доступа.

9. СТАТУСЫ ДОСТУПА К ПРОДУКТУ

Статус доступа должен фиксироваться для каждой страны, продукта, интерфейса и даты.

PA-0— НЕИЗВЕСТНО

Не удалось надежно определить статус доступа.

PA-1— НЕДОСТУПНО

Продукт не предлагается в соответствующей стране или группе пользователей.

PA-2— ОГРАНИЧЕННО

Продукт ограничен определенными:

- аккаунт,
- приглашениями,
- учебное заведение,
- не подразумевает никакой пригодности в отношении возраста,
- план,
- устройства,

условиями.

PA-3— ДОСТУПНО, НО НЕ ПОДДЕРЖИВАЕТСЯ ОФИЦИАЛЬНО

Фактический доступ к продукту может быть наблюдаем. Однако официальная поддержка или статус использования неясны или ограничены. Эти пользователи не должны автоматически добавляться в основную официальную панель населения. Можно создать отдельную экспериментальную панель.

PA-4— ОФИЦИАЛЬНО ДОСТУПНО

Продукт официально доступен, и основные условия использования определены.

PA-5— ОФИЦИАЛЬНО ДОСТУПНО С ПОЛНЫМ АУДИТОРСКИМ НАБОРОМ

Продукт не только недоступен; он контролируется:

- языка,
- план,
- интерфейс,
- в условиях сессии

подходит для генерации наблюдения, требуемого протоколом. Наличие продукта в стране не означает, что все планы и языки соответствуют контролю.

10. ОБХОД ОГРАНИЧЕНИЙ НЕ ЯВЛЯЕТСЯ РЕАЛЬНЫМ ИЗМЕРЕНИЕМ НАСЕЛЕНИЯ

Участникам нельзя просить выполнять следующие действия:

- Укажите неправильную страну
- Создайте фальшивый аккаунт
- Измените информацию о возрасте
- Обойти официальные ограничения доступа
- Используйте чужой аккаунт
- Обойдите защиту или проверку продукта
- Нарушение условий использования

Такое поведение также может рассматриваться как технический эксперимент. Однако:

Опыт естественной и легитимной пользовательской аудитории

не может быть представлен как таковой. Если в стране нет доступа, правильный статус должен быть: НЕДОСТУПНО / или / ВНЕ ДОСТУПА ДЛЯ НАТИВНОЙ АУДИТОРИИ. Одно наблюдение, превышающее ограничение, не формирует показатель населения страны.

11. РАЗЛИЧИЕ МЕЖДУ ПОТЕНЦИАЛЬНЫМ И АКТИВНЫМ ПОЛЬЗОВАТЕЛЕМ

Люди, которые могут использовать продукт, не являются тем же сообществом, что и люди, которые фактически используют продукт.

11.1. Потенциальная доступная популяция

Это люди, которые технически и законно могут использовать продукт. Эта популяция отвечает на вопрос: «Что увидел бы подходящий человек, если бы использовал продукт?» Эта популяция ценна для справедливости и сравнения продуктов.

11.2. Активные пользователи

Это люди, которые фактически использовали продукт в течение определенного периода. Эта популяция отвечает на вопрос: «Что видят люди в пределах фактического объема использования?» Для этого необходимы надежные данные о активном использовании.

11.3. Пользователи, подвергшиеся запросам

Это люди, которые используют продукт и имеют потенциальную возможность задавать вопросы о контролируемом бренде или теме. Эту популяцию часто нельзя определить напрямую. Могут потребоваться данные о поведении при поиске или запросах.

11.4. Популяция, фактически видящая ответ

Они являются пользователями, которые фактически подвергаются воздействию конкретного представленного вывода. Это может быть не полностью измеримо без данных реального поставщика продукта за пределами пользовательской панели.

12. ВЕС ВОЗДЕЙСТВИЯ НЕ ТАКОЙ ЖЕ, КАК ВЕС ПРАВОМОЧНОСТИ

Можно получить два отдельных глобальных результата.

12.1. Результат с взвешенной по популяции право на участие

Каждый пользователь взвешивается в соответствии с его долей в определенной популяции правомочных пользователей. Этот результат:

- глобально справедлив,
- потенциально доступен,
- представителен для популяции

в терминах

12.2. Результат с взвешенной по реальному воздействию

Продукты ИИ или группы пользователей взвешиваются в соответствии с фактическими объемами использования. Этот результат:

- ценен с точки зрения
- текущего коммерческого влияния,
- фактического воздействия на пользователей,

операционный риск.

Однако надежные данные об использовании могут быть недоступны. Оценки могут основываться на заявлениях поставщиков, рыночных данных или неполном выборочном обследовании. Следовательно, тип взвешивания должен быть явно указан.

12.3. Замена одного на другой запрещена

Баллы для допустимого населения: не означает «Большинство фактических пользователей видели этот ответ». Баллы за охват: не означает «Обеспечено справедливое представление мирового населения».

13. ОБЛАСТЬ ДОСТУПА И ТОЧНОСТЬ ПРЕДСТАВЛЕНИЯ ДОЛЖНЫ БЫТЬ РАЗДЕЛЕНЫ

Продукт ИИ может быть доступен только в ограниченном числе стран. В этих странах он может создавать очень точные представления. Другой продукт может иметь более широкий доступ, но давать менее точные результаты. Если рассматривать только точность представления, продукт с ограниченным доступом может казаться более превосходным. Если рассматривать только охват доступа, некорректное представление может остаться незамеченным. Поэтому нужны два отдельных результата.

13.1. Доступное покрытие

Указывает, какая часть указанной глобальной эталонной популяции может надлежащим образом использовать соответствующий продукт ИИ. Кандидат на представление:

```
AC_a= N(U_{a,t}) / N(G_t)
```

Здесь:

- Ua,t: допущенная пользовательская популяция продукта
- Gt: определенная глобальная эталонная популяция

13.2. Условная точность представления

Это вероятность увидеть точное представление среди пользователей, имеющих доступ к продукту. Кандидатское обозначение:

```
RA_a= Pr(Y=1| u ∈ U_{a,t})
```

13.3. Эффективный доступ к точному представлению

КАНДИДАТСКАЯ КОНЦЕПЦИЯ

Можно разработать операционный показатель, который сочетает охват доступа с условной точностью:

```
ECR_a= AC_a× RA_a
```

Эта метрика примерно соответствует вопросу: «Какая доля определённой глобальной эталонной популяции может получить доступ к этому продукту и выполнить условие получения точного представления?» Это ещё не окончательная формула NOMOS. Потому что:

- доступ отличается от фактического использования,
- структуры веса и зависимости сложны,
- использование продукта добровольно,

Правильное представление зависит от запроса. Однако оно сохраняет основное различие:

> *Неспособность получить доступ не равнозначна получению неправильного ответа.*

14. Естественный охват И Общая опора сравнения

При сравнении ИИ-продуктов требуются две отдельные популяционные вселенные.

14.1. Естественный охват Популяция

Каждый продукт ИИ имеет свою допустимую популяцию в рамках официального и фактического доступа. Этот результат отвечает на следующий вопрос:

> *В какой степени продукт точно отражает население, к которому он фактически может получить доступ?*

Целевые группы населения могут различаться в зависимости от продукта.

14.2. Общая опора сравнения Популяция

В общей форме для всех сравниваемых ИИ-продуктов:

- официально доступно,
- может использоваться при тех же условиях языка или интерфейса

это население. Этот результат отвечает на следующий вопрос:

> *Какой продукт обеспечивает более точное представление в той же пользовательской среде?*

14.3. Два результата должны быть опубликованы вместе

Естественный охват:

- показывает фактический доступ к продукту,

- естественное географическое распределение

Общая опора сравнения: предоставляет более справедливое сравнение продуктов. Продукт может иметь высокую оценку по Естественный охват, но ниже по Общая опора сравнения. Два результата отвечают на разные вопросы.

15. НАСЕЛЕНИЕ, СВЯЗАННОЕ С РЫНКОМ

Глобальный идентификационный показатель сущности может отличаться от показателя ее целевого коммерческого рынка. Пример: если местная юридическая фирма предоставляет услуги только в Турции: вопрос глобальной идентичности может измеряться среди всей допустимой пользовательской аудитории в мире, но рекомендации и пригодность услуг следует оценивать только для пользователей в Турции или в пределах фактического объема предоставляемых услуг. Учреждение может выбрать результат в свою пользу:

- только высоко оцениваемый рынок,
- только на своем языке,
- только своих существующих клиентов

целевая аудитория не может быть объявлена. Целевой рынок:

- Перед началом аудита,
- по данным о предоставленных услугах,
- в рамках контракта и операций

Он должен быть зафиксирован соответственно.

16. РАЗДЕЛЕНИЕ ЛИЦ, АККАУНТОВ И СЕССИЙ

Вес населения относится к людям. Не к AI-аккаунтам.

16.1. Один человек, несколько аккаунтов

Один и тот же человек:

- бесплатный,
- платный,
- корпоративный

может иметь несколько аккаунтов. Эти аккаунты могут давать отдельные наблюдения в сравнениях интерфейс продукта. Их нельзя учитывать как отдельных людей в глобальной человеческой популяции.

16.2. Один аккаунт, несколько человек

Семейный, командный или корпоративный аккаунт может использоваться несколькими людьми. Это не равно количеству пользователей аккаунта.

16.3. Один человек, несколько сессий

Человек может создавать отдельные наблюдения в разных сессиях. Эти наблюдения могут измерять время или стабильность сессий. Они не увеличивают размер независимой популяции.

16.4. Один человек, несколько продуктов ИИ

Это ценно для сравнения продуктов в паре. Однако наблюдения зависимы из-за общих характеристик на уровне пользователя.

16.5. Основная единица веса

В оценке GEO на основе популяции основная единица веса должна быть:

человек

Расчеты, сессии и наблюдения за продуктом должны быть связаны с единицей человека.

17. ЯЗЫК ПОПУЛЯЦИИ

Использование только официального языка страны недостаточно для языковой справедливости.

17.1. Язык промпта

Язык, на котором участник задает вопрос аудита.

17.2. Язык интерфейса

Язык, используемый в меню и пользовательском интерфейсе ИИ-продукта. Он может отличаться от языка промпта.

17.3. Язык ответа

Язык, на котором ИИ фактически отвечает. Система может отвечать на другом языке, отличном от языка запроса. Это также является результатом измерения.

17.4. Язык владения пользователем

Язык, на котором участник может осмысленно использовать запрос и ответ. Это не обязательно должен быть родной язык.

17.5. Локаль

Региональный контекст вместе с языком. Один и тот же язык может быть разным:

- страны,
- цена,
- закон,
- терминология

может влиять на условие.

18. ДВОЙНОЙ УЧЕТ МНОГОЯЗЫЧНЫХ ПОЛЬЗОВАТЕЛЕЙ

Человек:

- Турецкий,
- может владеть английским языком.
- Немецкий

Этот человек может пройти три отдельных языковых теста. Однако учет его как трех человек в глобальном расчете населения искажает оценку численности. Можно использовать три метода.

18.1. Задание по родному языку для одинокой матери

Человек назначен на одну языковую ячейку в основном анализе. Другие языки используются в качестве дополнительных или парных тестов.

18.2. Разделить вес человека

Вес населения человека можно распределить между соответствующими языками в заранее определённой манере. Общий вес не превышает одного человека.

18.3. Отдельная многоязычная согласованная панель

Многоязычные пользователи используются для сравнений одного и того же человека на разных языках. Эта панель анализируется отдельно от основной панели населения. Независимо от используемого метода:

> *Общий вес населения одного и того же человека не может быть воспроизведён без объяснения.*

19. КАК ПРОВЕРИТЬ ВЛАДЕНИЕ ЯЗЫКОМ?

Заявление участника: «Я знаю этот язык.» может быть первоначальной записью. Однако при аудите с высокой степенью безопасности может использоваться один из следующих методов:

- Декларация родного языка или ежедневного использования
- Краткий тест на чтение и понимание
- Языковая запись поставщика панели
- Проверка местным арбитром
- Запись о школьном или профессиональном использовании
- Успешность пилотной миссии

Панели тестов на чрезвычайно трудные языки могут быть доступны подготовленным и высокодоходным пользователям. Цель не в проведении академического языкового экзамена. Цель — проверить способность понимать запрос и ответ как реальный пользователь.

20. ЧТО ПРОИСХОДИТ, ЕСЛИ ЯЗЫК НЕ ПОДДЕРЖИВАЕТСЯ?

Продукт ИИ может:

- не поддерживать конкретный язык официально,
- частично поддерживать его,
- не заявлять о поддержке, но уметь генерировать ответы

Возможны такие случаи. Статусы языков следует регистрировать отдельно:

ОФИЦИАЛЬНО_ПОДДЕРЖИВАЕТСЯ

ОГРАНИЧЕННАЯ_ПОДДЕРЖКА

НАБЛЮДАЕМО, НО НЕ ПОДДЕРЖИВАЕТСЯ

НЕ ПОДДЕРЖИВАЕТСЯ

НЕИЗВЕСТНО

Ответ получен на языке, который официально не поддерживается:

- обнаружение,
- экспериментальный доступ,
- устойчивость языка

может быть измерена. Не следует автоматически добавлять в основной балл поддержки пользователей. Также должен быть зафиксирован предел доступа к языку.

21. ВЗВЕШИВАНИЕ НАСЕЛЕНИЯ СТРАНЫ И ОХВАТ НАБЛЮДАТЕЛЯ СТРАНЫ

Основная панель по взвешенному населению и панель видимости по странам служат разным целям.

21.1. Панель населения

Участники распределяются в соответствии с их долей в целевой аудитории, подходящей для продукта. Более крупные допустимые популяции получают больший вес. Производит основной глобальный балл по населению.

21.2. Панель наблюдателей страны

Панель наблюдателей стран стремится получить как минимум одно наблюдение от более мелких стран, которые получают мало или совсем не получают представления в основной выборке населения. Одно наблюдение Панели наблюдателей страны:

- означает, что доступ имеется,
- что произошло конкретное единичное событие,
- показывает необходимость нового исследования

может указывать. Сама по себе не производит оценку страны.

21.3. Достаточность для оценки страны

Чтобы можно было публиковать показатели по странам:

- минимальный образец,
- должна быть действительная норма наблюдений,
- неопределенность,
- репрезентативность панели

должны быть выполнены условия. При наличии одного или нескольких участников: 'Оценка страны NOMOS' не может быть объявлена.

22. Панель языковой справедливости

Вес населения может сделать языки с низкой численностью населения или ограниченными ресурсами невидимыми в основном рейтинге. Поэтому может быть создана отдельная Панель языковой справедливости. Цель:

- тестировать недопредставленные языки,
- обнаруживать серьёзное ухудшение качества на основе языка,
- выявлять проблемы с переводом промптов,
- делать поведение языков с ограниченными ресурсами видимым

должно быть. Панель языковой справедливости может предоставлять дополнительные образцы основной панели населения. Однако наблюдения за переизбыточно представляемыми языками должны возвращаться к фактическим весам населения в глобальном рейтинге. В противном случае панель справедливости изменяет оценку глобальной численности населения.

23. ПАНЕЛЬ ДОСТУПНОСТИ

Пользователи с визуальными, слуховыми, моторными, когнитивными или другими потребностями в доступности являются реальной частью целевой аудитории. Если задача NOMOS Capture или исследовательская задача не подходит для участия этих людей, пользователей нельзя исключать из целевой аудитории. Правильный результат:

> *Рамка выборки неполна с точки зрения доступности.*

Отдельная панель по доступности может оценить следующее:

- Совместимость с экранными считывателями
- Использование клавиатуры
- Голосовой ввод
- Настройки для слабовидящих
- Познавательная нагрузка
- Доступность на мобильных устройствах
- Удобство использования инструмента Capture

Панель по доступности должна рассматривать пользователей как часть охвата аудитории, а не как "особое исключение".

24. ЦЕЛЕВАЯ АУДИТОРИЯ И РАМКА ВЫБОРКИ

Эти две концепции не следует путать друг с другом.

24.1. Целевая популяция

Вся группа людей, о которой хотят сделать вывод.

24.2. Рамки выборки

Популяция, из которой участники могут быть фактически отобраны:

- панель,
- список,

- база данных,
- совместная организация,
- пул пользователей

является кластером. Рамки выборки могут не полностью охватывать целевую популяцию.

24.3. Контактируемая популяция

Это люди, которым исследовательское приглашение может быть фактически доставлено.

24.4. Популяция, согласившаяся участвовать

Те, кто добровольно участвует в исследовании.

24.5. Популяция, выполняющая задание

Те, кто выполняет задачу аудита.

24.6. Действительная аналитическая популяция

Это те, кто производят наблюдения, соответствующие протоколу и пригодные для анализа. Числа этих слоев должны быть указаны отдельно.

25. ОШИБКА ПОКРЫТИЯ

Выборочная совокупность может исключать некоторые части целевой популяции. Это называется ошибкой покрытия. Примеры:

- Только профессиональные онлайн-панели
- Только пользователи настольных компьютеров
- Только люди с кредитными картами
- Только те, кто использует английский интерфейс
- Только жители крупных городов
- Только пользователи, сильно интересующиеся технологиями
- Исключение людей с потребностями в доступности

Ошибка покрытия:

- должна быть объяснена,
- должна быть измерена, если это возможно,

следует сократить с помощью взвешивания или дополнительной панели. Взвешивание не может волшебным образом создать группу, которой нет в выборочной рамке.

26. ПЕРЕОХВАТ

Люди, которые не принадлежат к целевой популяции, могут встречаться в выборочной рамке. Примеры:

- Участник из страны, где продукт недоступен
- Лицо, не соответствующее требуемому возрастному условию
- Неправильный план или пользовательский интерфейс
- Участник, который не может использовать язык аудита
- Повторное обнаружение лица с несколькими регистрациями

- Пользователь, работающий от имени другой страны

Записи с избыточным охватом не следует включать в основной показатель. Однако причина исключения и процент должны быть зафиксированы.

27. НЕОТВЕТЫ И САМОИЗБОР

Люди, согласившиеся участвовать в исследовании, могут систематически отличаться от целевой популяции. Волонтеры:

- Могут чаще использовать продукты с ИИ,
- Могут быть более заинтересованы в технологиях,
- Могут иметь более высокий уровень образования,
- Могут уже быть знакомы с брендом,

Могут вести себя иначе из-за вознаграждения или компенсации. Следовательно, выражение "1,000 случайные люди в мире" следует использовать только в действительно вероятностных и документированных дизайнах отбора. Если используется волонтер или онлайн-панель, правильное выражение:

Слоистая и взвешенная панель добровольных пользователей

может быть возможна. Тип панели не может быть сохранен.

28. ПОДХОДЯЩАЯ ПОПУЛЯЦИЯ МЕНЯЕТСЯ СО ВРЕМЕНЕМ

Всемирная база пользователей продукта с ИИ не является фиксированной. Могут изменяться следующие параметры:

- Поддерживаемые страны
- Минимальный возраст
- Бесплатный доступ
- Структура платного плана
- Доступность на мобильных устройствах или в вебе
- Поддержка языков
- Условия учетной записи
- Охват продукта для корпоративных клиентов
- Юридический и нормативный статус

По этой причине каждая запись о населении:

- дата наблюдения,
- период действия,
- версия источника,
- дата последней проверки

должны содержать.

28.1. Изменения во время волны измерений

Если доступ к продукту существенно изменяется во время волны:

- волну можно остановить,

- можно создать два отдельных периода до и после изменения,

затронутые страны можно анализировать отдельно. Пользователи до и после изменения не могут быть объединены в одно население.

29. ИСХОДНЫЙ ФАЙЛ НАСЕЛЕНИЯ

Каждый аудит должен хранить используемые целевые данные в версионном файле. В зависимости от этого следует включать следующие поля:

- Страна
- Эталонная популяция
- История населения
- Возрастная возможность участия
- Интернет и доступ к устройствам
- Статус доступа к продукту
- Доступ к интерфейсу
- Требования к плану и счету
- Языковые группы
- Право на участие в исследованиях
- Распространённый метод распределения
- Неопределенность
- Источники
- Дата последней проверки
- Ответственный человек
- Версия численности населения

Квоты по странам не могут быть воспроизведены после результатов без этого файла.

30. СИНТЕТИЧЕСКИЙ ПРОДУКТ И ЭКСПОЗИЦИЯ СТРАНЫ

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Следующие страны, продукт ИИ, численность населения и показатели доступа полностью вымышлены. Они не представляют ни одну реальную страну или продукт. Рассмотрим синтетический общий потребительский продукт под названием Orion AI. Существуют четыре синтетические страны.

Страна	Общая численность населения	Примерная подходящая часть населения с учетом возраста, интернета, языка и условий интерфейса	Доступ к продукту	Население в пределах естественного охвата
Страна A	1,400,000,000	780,000,000	Недоступно	0
Страна B	1,450,000,000	520,000,000	Официальный доступ	520,000,000
Страна C	340,000,000	250,000,000	Официальный доступ	250,000,000
Страна D	35,000	25,000	Официальный доступ	25,000
Итого	3,190,035,000	1,550,025,000		770,025,000

30.1. Если использовать все население

Страна A имела бы большую долю в основной выборке из-за своего очень большого населения. Тем не менее, продукт не может использоваться в этой стране. Участники не могли бы получить естественный пользовательский опыт.

30.2. Когда используется население в пределах естественного охвата

Теоретические доли населения основной панели продуктов составляют приблизительно:

- Страна B: 67.53%
- Страна C: 32.47%
- Страна D: 0.003%
- Страна A: за пределами Естественный охват

Теоретическая доля Страны D в основной выборке численностью 1,000 человек значительно меньше одного человека. Следовательно, основной показатель с учетом населения не должен сильно учитывать Страну D. Однако, если страна не предназначена для полного игнорирования:

Панель наблюдателей страны

Можно получить как минимум одно исследовательское наблюдение. Это наблюдение:

- не является показателем Страны D,
- не изменяет вес основного населения,

показывает, что поведение продукта можно наблюдать в стране.

30.3. Как отражается Страна A?

Для Страны A:

Используется статус OUTSIDE_NATIVE_REACH

исключение Страны A не должно скрывать ограничения доступа к продукту. Публичная карточка результатов также показывает:

- Покрытие доступа к продукту
- Доля, недоступная в глобальной эталонной популяции
- Оценка Естественный охват
- Сравнительная оценка Общая опора сравнения

31. ПРЕДСТАВЛЕНИЕ СИНТЕТИЧЕСКОЙ ПОПУЛЯЦИИ ЗАПРОСОВ

СИНТЕТИЧЕСКИЙ ПРИМЕР

Рассмотрим три запроса для одного и того же продукта Orion AI.

Запрос A

"Что за компания Asteron?" Целевая популяция: может быть широкой подходящей пользовательской популяцией Orion AI.

Запрос B

"Предоставляет ли Asteron услуги по корпоративным поездкам в Турции?" Целевая популяция:

- Подходящие пользователи в Турции,
- Определены внешние пользователи, изучающие сервис Турции

может быть.

промпт C

«Подходит ли Asteron для консультаций по легальной иммиграции для меня?» Если Asteron не предоставляет такую услугу: возможно, нет населения, имеющего право на консультацию, правильный ответ должен исключать всех подходящих пользователей. Объединение этой промпта в общий показатель идентичности может быть неверным. Этот пример показывает:

> *Даже для одной и той же сущности и одного и того же продукта ИИ, целевая аудитория может различаться в зависимости от измеряемого вопроса пользователя.*

32. АРХИТЕКТУРА ИЗ ЧЕТЫРЕХ ПАНЕЛЕЙ

АРХИТЕКТУРА КАНДИДАТА — АРХИТЕКТУРА КАНДИДАТА

Рекомендуются четыре взаимодополняющие панели для глобальной и справедливой реализации GEO-1000.

32.1. Панель населения

Основная панель с весом, соответствующим населению пользователей, имеющих право на продукт. Производит основную оценку населения.

32.2. Панель наблюдателей страны

Предотвращает полное исчезновение малых стран, которые не могут предоставить достаточное количество образцов в панельной выборке. Одного этого недостаточно, чтобы получить оценку страны.

32.3. Панель языковой справедливости

Также измеряет языки с низким ресурсом или подавляемые в средних показателях по населению. Он вновь соединяется с глобальной основной оценкой с учетом реальных весов населения.

32.4. Панель доступности

Измеряет потребности в доступности и пользователей вспомогательных технологий, которых стандартная цифровая панель может исключать. Наблюдения из этих четырех панелей:

- цель,
- вес,
- модель вывода

Это следует маркировать отдельно с точки зрения обслуживания. Одно и то же наблюдение не может использоваться с полной весовой нагрузкой более чем в одной панели без объяснения.

33. ПАНЕЛЬ ДЛЯ ВЗРОСЛЫХ (НАЧАЛЬНАЯ)

ПРАВИЛО КАНДИДАТА — ОСНОВНОЕ ПРАВИЛО КАНДИДАТА

В первом общем выпуске GEO-1000, приборные панели реальных пользователей:

от взрослых участников, которые соответствуют условиям продукта и страны

Рекомендуется создать. Этот подход:

- согласие,
- защита данных,
- собирая скриншоты,
- исследования с высоким риском,
- родительское согласие

Это может уменьшить его сложность. Однако следующий вывод сделать нельзя: 'Опыт детей и молодых пользователей был измерен с помощью взрослой панели.' В продуктах, которые затрагивают детей или молодежь, отдельно:

- этическое одобрение,
- протокол, соответствующий возрасту,
- процессы родителей или опекунов,
- защита от вреда

требуются.

34. НИКАКАЯ РАМКА НАСЕЛЕНИЯ НЕ МОЖЕТ НАЗЫВАТЬСЯ "ВСЕ ЧЕЛОВЕЧЕСТВО"

Продукт:

- если его нет в некоторых странах,
- если он не поддерживает некоторые языки
- если это работает только на определённых устройствах,
- если есть определённое условие по возрасту или тарифу,

результат: это нельзя представлять как «ГЕО-опыт всех людей в мире». Правильное выражение:

Оценочное представление в целевой популяции, подходящее для указанного продукта

должно использоваться. Слово глобальный:

- много стран,
- многоязычный,
- большое население

может определять измерение. Оно не означает неограниченное и универсальное покрытие.

35. ФУНДАМЕНТАЛЬНОЕ ЭТИЧЕСКОЕ ОГРАНИЧЕНИЕ НАСЕЛЕНИЯ

Определение популяции для измерения не означает, что некоторые люди считаются недостойными. Группа может получать низкий вес в основном балле. Однако:

- высокий риск,
- уазу справедливости,
- доступность,
- видимость страны

может быть также рассмотрено отдельно. Группа может быть исключена из показателя Естественный охват, потому что у неё нет доступа к продукту. Этот недостаток доступа также может быть зафиксирован как глобальный пробел покрытия. Цель NOMOS:

- сжать всех в один искусственный средний показатель,
- видеть только крупнейшие популяции,
- казаться, что символически представлять маленькие сообщества одним человеком

не является целью. Цель заключается в следующем:

> *Представить каждого человека в правильном знаменателе правильного вопроса, не смешивая его вес и видимость.*

36. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH05-N01

Целевая пользовательская популяция должна быть определена отдельно для каждого ИИ-продукта.

CH05-N02

Общая численность населения страны не может напрямую использоваться как население, имеющее право на AI.

CH05-N03

Целевая популяция должна быть связана с продуктом, страной, возрастом, доступом, интерфейсом, планом, языком, промптом и временными условиями.

CH05-N04

Целевая аудитория и правила исключений должны быть версионированы до просмотра результатов.

CH05-N05

Целевая аудитория ИИ-продукта не может быть перенесена на другой ИИ-продукт без объяснения.

CH05-N06

Отсутствие доступа к продукту не может оцениваться как результат искажения информации; оно должно сообщаться отдельно как охват доступа.

CH05-N07

Участников нельзя просить обойти ограничения на доступ к продукту или стране.

CH05-N08

Доступ, официально не поддерживаемый, не может быть автоматически добавлен в основную популяцию Естественный охват.

CH05-N09

Единицы индивидуальные, учетные и сеансовые должны быть отделены друг от друга.

CH05-N10

Несколько учетных записей или языков одного человека не могут умножить вес человека в населении.

CH05-N11

Право на использование языка не может присваиваться только на основании официального языка страны.

CH05-N12

Запрашиваемый язык, язык интерфейса, язык ответа и язык владения пользователя должны быть записаны отдельно.

CH05-N13

Общий вес населения человека, использующего несколько языков, не может превышать вес одного человека без объяснения.

CH05-N14

Панель с взвешиванием по населению и панель Country Observer должны оставаться отдельными.

CH05-N15

Одно или несколько наблюдений Country Observer не могут публиковаться как оценка страны.

CH05-N16

В панели Language Fairness наблюдения с избыточной выборкой должны быть приведены к их истинному весу в глобальной оценке.

CH05-N17

Пользователи с особыми потребностями не могут быть исключены из целевой популяции из-за того, что инструмент сбора данных не подходит.

CH05-N18

Целевая популяция, рамка выборки, контактируемая популяция, участники и достоверные аналитические наблюдения должны быть представлены отдельно.

CH05-N19

Неправдоподобная или добровольная панель не должна представляться как случайная выборка глобального населения.

CH05-N20

Взвешивание не может сделать наблюдаемой группу, полностью отсутствующую в рамке выборки.

CH05-N21

Потенциальная подходящая популяция и воздействие со стороны активных пользователей не могут использоваться как одна и та же метрика.

CH05-N22

Результаты Естественный охват и Общая опора сравнения должны быть разделены друг от друга.

CH05-N23

Одна и та же целевая популяция не может автоматически использоваться для идентификации, услуг, цены, юридических и рекомендательных запросов.

CH05-N24

Данные о популяции, статус доступа к продукту и распределение по языкам должны содержать временную метку и версию источника.

CH05-N25

Коэффициенты предельной полезности нельзя слепо умножать без объяснения предположения об их независимости.

CH05-N26

Если тип продукта не подходит для общей панели потребителей, метод исследования человеческой популяции GEO-1000 не может применяться без объяснения.

CH05-N27

Дети или несовершеннолетние пользователи не могут включаться в основную панель без отдельного этического и юридического протокола.

CH05-N28

Охват доступа к продукту и точность представительства должны публиковаться как отдельные результаты.

CH05-N29

Термин «глобальный» не может сделать невидимыми страны, языки, возраст, интерфейсы или планы, которые выходят за рамки.

CH05-N30

Запись о целевой популяции должна иметь ответственного человека или институционального владельца.

37. ФОРМЫ ОШИБОК

CH05-F01— ПОДСЧИТЫВАЙТЕ ОБЩУЮ ПОПУЛЯЦИЮ КАК ПОДХОДЯЩУЮ ПОПУЛЯЦИЮ

Население страны включается в знаменатель напрямую без оценки условий продукта, возраста, интернета, языка и интерфейса.

CH05-F02— ПРИМЕНЯЙТЕ ОДНУ И ТУ ЖЕ КАРТУ МИРА КО ВСЕМ ПРОДУКТАМ

Продукты ИИ с разными зонами доступа измеряются по одной и той же целевой популяции.

CH05-F03— ПОДСЧИТЫВАЙТЕ ПОЛЬЗОВАТЕЛЯ, ПРЕВЫШАЮЩЕГО ЛИМИТ, КАК ЕСТЕСТВЕННОГО ПОЛЬЗОВАТЕЛЯ

Наблюдения, превышающие условия официального доступа, представлены как опыт населения страны.

CH05-F04— ПОДСЧЁТ ОТСУТСТВИЯ ПРОДУКТА КАК НЕПРАВИЛЬНОГО ОТВЕТА

Страны, которые недоступны, считаются нулевыми по точности представления.

CH05-F05— ПОЛНОЕ СКРЫТИЕ ОТСУТСТВИЯ ПРОДУКТА

Большие недоступные группы населения исключаются из знаменателя Естественный охват; охват доступа также не отображается.

CH05-F06— УЧЕТ ОДНОГО ЧЕЛОВЕКА БОЛЕЕ ОДНОГО РАЗА

Аккаунт одного и того же человека, язык или наблюдения за продуктом умножают вес населения.

CH05-F07— УЧЕТ АККАУНТА КАК ЧЕЛОВЕКА

Множественное владение аккаунтами или совместное использование аккаунта не учитывается.

CH05-F08— УЧЁТ ОФИЦИАЛЬНОГО ЯЗЫКА КАК ЯЗЫКА ПОЛЬЗОВАТЕЛЯ

Официальный язык страны считается языком аудита для всех участников.

CH05-F09— ДВОЙНОЙ УЧЕТ МНОГОЯЗЫЧНЫХ ЛИЦ

Один и тот же человек получает полный вес населения в каждом языке.

CH05-F10— ПРИСВОЕНИЕ ОДНОГО ЧЕЛОВЕКА МАЛОЙ СТРАНЕ И ОБЪЯСНЕНИЕ ОЦЕНКИ СТРАНЫ

Наблюдение за страной наблюдателем преобразуется в статистическую оценку страны.

CH05-F11— ПОЛНОЕ УДАЛЕНИЕ МАЛЫХ СТРАН

Поскольку доля населения невелика, страна или язык не включаются ни в какую дополнительную панель покрытия.

CH05-F12— ПРЕВЫШЕНИЕ ВЕСА ПРИМЕРОВ СПРАВЕДЛИВОСТИ ЯЗЫКА В ОСНОВНОЙ ОЦЕНКЕ

Языки с чрезмерной выборкой не возвращаются к их истинному весу в населении.

CH05-F13— УЧЕТ АКТИВНОГО ПОЛЬЗОВАТЕЛЯ КАК ПОТЕНЦИАЛЬНОГО НАСЕЛЕНИЯ

Фактические данные об использовании представлены так, как если бы они включали всех людей, которые могли бы получить доступ к продукту.

CH05-F14— УЧЕТ ПОТЕНЦИАЛЬНОГО НАСЕЛЕНИЯ КАК ФАКТИЧЕСКОГО ВОЗДЕЙСТВИЯ

Предполагается, что ответ каждого, кто мог использовать продукт, на самом деле наблюдается.

CH05-F15— УЧИТЫВАНИЕ ДОБРОВОЛЬНОЙ ПАНЕЛИ КАК СЛУЧАЙНОЙ ВЫБОРКИ МИРОВОГО НАСЕЛЕНИЯ

Самовыбор и пристрастие панели скрыты.

CH05-F16— ИГНОРИРОВАНИЕ ОШИБКИ ПОКРЫТИЯ ПРИ ВЕСОВОЙ КОРРЕКТИРОВКЕ

Утверждается, что группы, отсутствующие в рамке, представлены исключительно посредством взвешивания.

CH05-F17— ИСКЛЮЧЕНИЕ ИЗ НАСЕЛЕНИЯ ПОЛЬЗОВАТЕЛЕЙ, ОТСЕЯННЫХ ИНСТРУМЕНТОМ СБОРА ДАННЫХ

Проблемы с доступностью или устройством интерпретируются как несоответствие пользователя.

CH05-F18— ПРИМЕНЕНИЕ ОДНОГО И ТОГО ЖЕ НАСЕЛЕНИЯ КО ВСЕМ промптам

Вопросы глобальной идентичности, локального ценообразования и юридические вопросы объединены в одном знаменателе.

CH05-F19— ПРИМЕНЕНИЕ ОБЩЕГО ПОТРЕБИТЕЛЬСКОГО НАСЕЛЕНИЯ К API-ПРОДУКТУ

Панель пользователей передается напрямую разработчику или продукту приложения.

CH05-F20— СЕКРЕТНОЕ СОЧЕТАНИЕ БЕСПЛАТНЫХ И ПЛАТНЫХ ПЛАНОВ

Одиночное распространение продукта публикуется без фиксирования различий в плане.

CH05-F21— ОБЪЕДИНЕНИЕ ВЕБ-И МОБИЛЬНЫХ ПОПУЛЯЦИЙ

Различия в доступе к интерфейсу и поведении остаются незаметными.

CH05-F22— ИГНОРИРОВАНИЕ ВОЗРАСТНОГО ЛИМИТА

Неподходящие возрастные группы добавляются в целевую популяцию или панель без объяснения.

CH05-F23— УМНОЖЕНИЕ ПРЕДЕЛЬНЫХ СТАВОК КАК БУДТО НЕЗАВИСИМЫХ

Игнорируется связь между возрастом, интернетом, языком и доступом.

CH05-F24— ИЗМЕНЕНИЕ ПОПУЛЯЦИИ ПОСЛЕ РЕЗУЛЬТАТА

Низкооценённая страна или группа пользователей удаляются из целевой популяции.

CH05-F25— ИСПОЛЬЗОВАНИЕ СТАРОЙ КАРТЫ ДОСТУПА ПРОДУКТА

Фактический охват продукта на дату измерения не проверяется.

CH05-F26— ПУТАНИЦА В РЕЗУЛЬТАТАХ Общая опора сравнения И Естественный охват

Сравнения, сделанные на совместной популяции, представлены так, как если бы они отражали фактический глобальный охват продукта.

CH05-F27— СУЖЕНИЕ ЦЕЛЕВОГО РЫНКА В СООТВЕТСТВИИ С КОММЕРЧЕСКИМИ ИНТЕРЕСАМИ

Организация объявляет целевыми рынками только те страны, в которых она добивается успеха.

CH05-F28— ПРИНЯТИЕ ОБЪЕМА ДОСТУПА ЗА ГЕОГРАФИЧЕСКУЮ ТОЧНОСТЬ

Присутствие во многих странах представлено как доказательство точного отображения.

CH05-F29— ПРЕДСТАВЛЕНИЕ ОГРАНИЧЕННОГО ДОСТУПА КАК ВЫСОКОГО ГЕОГРАФИЧЕСКОГО УСПЕХА

Высокая точность в небольшом количестве стран представлена как глобальное преимущество.

CH05-F30— ИСПОЛЬЗОВАНИЕ ТЕРМИНА «ВСЕ МИРОВОЕ СОБЫТИЕ» БЕЗ ЗАПИСИ ПОКРЫТИЯ

Исключения продукта, языка, страны и плана скрыты.

38. ПРОЦЕДУРА АУДИТА

Шаг 1— Зафиксировать продукт ИИ и пользовательский интерфейс

План интерфейса продукта, тип учетной записи, отображаемая модель, дата измерения зафиксированы.

Шаг 2— Классифицировать вопрос измерения

промпт:

- глобальная идентичность,
- локальная услуга,
- цена,

- закон,
- рекомендация,
- высокий риск

к какой из их категорий это относится?

Шаг 3— Определить глобальную эталонную популяцию

Определен первоначальный человеческий универсум, к которому будет относиться результат.

Шаг 4— Определить возраст и соответствие критериям исследования

Возрастной статус продукта, условия исследования по возрасту, согласие на обработку данных и условия обработки данных зафиксированы.

Этап 5— Проверка доступа к продукту в зависимости от страны

Для каждой страны присваивается статус доступа между PA-0 и PA-5.

Этап 6— Определение совместимости с интернетом, устройствами и интерфейсами

Оценивается популяция, которая может использовать измеряемую интерфейс.

Этап 7— Установление языковой и локальной среды

Разделяются язык запроса, язык интерфейса, язык ответа и уровень владения языком пользователя.

Этап 8— Применение релевантности запроса

Оценивается, принадлежат ли все пользователи к целевой популяции соответствующей семейства запросов.

Этап 9— Выбор модели популяции, соответствующей типу продукта

Разделяются потребительский, корпоративный, образовательный, API или встроенный продукт.

Шаг 10— Сравните целевую популяцию с выборочной рамкой

Это фиксируется как недоохват и переохват.

Шаг 11— Определите основные и дополнительные панели

Объемы Панели населения, Панели наблюдателей стран, Панели языковой справедливости и Панели доступности имеют версии.

Шаг 12— Разделите Универсумы Естественный охват и Общая опора сравнения

Результаты внутри продукта и популяции для межпродуктового сравнения фиксируются отдельно.

Шаг 13— Разделите объем доступа и результаты точности

Точность между популяцией, которую продукт может достичь, и пользователями, достигнутыми продуктом, сообщается отдельно.

Шаг 14— Зафиксируйте файл популяции

Источники данных, предположения, дата и версия популяции закрепляются.

Шаг 15— Отслеживайте существенные изменения

Если доступ к продукту или условия меняются в процессе измерения, создается новая версия.

39. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Идентичность ИИ-продукта; пользовательский интерфейс; план и условия учетной записи; дата измерения; страны, в которых продукт доступен; статус доступа на уровне страны; возрастные условия; юридические и этические условия участия; записи по населению страны; возрастные распределения; записи об интернете и доступе к устройствам; распределение языков и локалей; поддержка языков продукта; доступность интерфейса; целевая группа промптов; определение целевой популяции; рамки выборки; недостаточное покрытие; избыточное покрытие; тип панели; метод отбора участников; вероятность или статус добровольного участия в панели; коэффициенты для многоязычных пользователей; покрытие Country Observer; покрытие Language Fairness; покрытие Accessibility Panel; популяция в пределах естественного охвата; Население Общая опора сравнения; данные активных пользователей, где использовались; расчет охвата доступа; предположения и анализ чувствительности; версия населения; и дата последней проверки.

Ответственное лицо или организация

40. ЧЕК-ЛИСТ АУДИТА

Какой продукт ИИ и пользовательский интерфейс измеряются? На какую дату была определена целевая популяция? Были ли отделены общая численность населения страны и численность подходящей пользователей? Был ли применён возрастной критерий? Оценивался ли доступ к интернету и устройствам? Доступен ли продукт официально по странам? Просили ли участников преодолевать ограничения доступа? Были ли разделены бесплатные и платные планы? Было ли разделение веб- и мобильных интерфейсов? Изменяла ли семья промптов целевую популяцию? Как определялась пригодность языка и локали? Учитывались ли многоязычные пользователи дважды? Ясна ли разница между целевой популяцией и выборочной совокупностью? Является ли панель вероятностной или добровольной? Измерялась ли ошибка покрытия? Пользователи с особыми потребностями в доступности были исключены без уведомления?

Были ли наблюдения Country Observer преобразованы в оценку страны? Имели ли примеры Language Fairness чрезмерное влияние на основную оценку? Были ли результаты Естественный охват и Общая опора сравнения разделены? Являются ли охват и точность представления отдельными показателями? Были ли результаты активных пользователей и потенциальных пользователей разделены? Подходит ли тип продукта для панели человеческой популяции? Были ли предельные ставки в весах популяции просто умножены? Было ли определение популяции изменено после просмотра результатов? Используются ли версии данных о популяции и доступе? Видны ли нетестированные группы? Соответствует ли термин «глобальный» фактическому охвату? Есть ли четко назначенный ответственный за запись популяции?

41. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Разве не будет справедливее распределять напрямую в соответствии с численностью населения мира?»

Мировое население является сильной исходной базой. Однако включение людей, которые не могут использовать продукт, в основную выборку продукта не дает реального пользовательского опыта. Правильным методом является:

- вес населения в пределах населения, имеющего право на продукт,
- глобальное покрытие доступа.

Оба результата должны быть представлены.

Возражение 2— «Разве большинство пользователей не должны быть из крупных стран?»

Страны, которые соответствуют условиям продукта, языка и доступа, естественно, должны иметь высокую долю. Однако только потому, что общая численность населения велика, невозможно создать поддельный опыт пользователя из страны, где продукт недоступен.

Возражение 3— «Почему было бы неправильно дать одного человека такой маленькой стране, как Сан-Марино?»

Человек ценен для исследования и наблюдения за страной с помощью Country Observer. Неправильно — представлять результат этого человека как "оценку страны". Видимость маленькой страны можно защитить с помощью панели Country Observer. Основной вес населения остается в соответствии с фактической долей населения.

Возражение 4— «Если мы дадим хотя бы одного человека каждой стране, разве мы не будем выглядеть более глобально?»

Предоставляется более широкое географическое покрытие. Однако оценка численности населения может быть искажена. Поэтому:

- географический охват,
- плотность населения

Это отдельные результаты. Это может быть панель охвата, чтобы выделить одного человека на каждую страну. Это не панель населения.

Возражение 5— «Если мы дадим ноль продукту в недоступной стране, не будет ли продукт несправедливо наказан?»

Недоступная страна не учитывается как ноль в точности представления. Она становится видимой в метрике охвата доступа. Это различие защищает продукт как от несправедливых штрафов, так и от ложной универсальности.

Возражение 6— «Если есть реальные данные об использовании, зачем нужны данные о населении?»

Реальные данные об использовании показывают операционное влияние. Однако продукт:

- какие населенные группы он не охватил,
- географическое неравенство в базе пользователей,
- Потенциальная пользовательская справедливость

Это не видно в одиночку. Оба типа данных отвечают на разные вопросы.

Возражение 7— «Если все пользователи будут задавать один и тот же запрос, почему важна релевантность запроса?»

Вопросы идентичности могут иметь значение для широкой аудитории. Вопросы по цене, юридическим вопросам или советам важны только для определённых пользователей и на рынках. Нерелевантная популяция может искажать распределение ответов от реальности пользователя.

Возражение 8— «Если человек знает несколько языков, почему его не должны полностью учитывать во всех языках?»

Все языки можно тестировать в межъязыковом сравнении. Однако многократный учёт одного и того же человека при взвешивании населения искусственно увеличивает численность мирового населения. Вес такого человека следует разделить либо использовать отдельную сопоставленную панель.

Возражение 9— «Разве тестирование только взрослых не оставляет глобальный результат неполным?»

Да, это так. Это ограничение должно быть чётко указано. Начало с взрослой панели может быть этическим и операционным выбором. Это не является суждением о детских пользователях.

Возражение 10— «Разве онлайн-панель для исследований не является достаточной?»

Она может быть очень полезной для пилотных исследований и исследований с взвешиванием. Но панель:

- охватывает, какая часть целевой аудитории,
- кто остаётся за пределами,
- как добровольное участие создаёт предвзятость

Это должно быть объяснено. Использование панели не является ошибкой. Ошибка — представлять панель как случайную, подобную мировой популяции.

Возражение 11— «Почему панель доступности связана с основным баллом GEO?»

Потому что часть реальных пользователей ИИ использует вспомогательные технологии или различные формы взаимодействия. Если инструмент сбора данных исключает этих людей, основной балл может не отражать всю целевую аудиторию.

Возражение 12— «Не делает ли наличие столь большого количества различных определений популяции невозможным получение одного единственного NOMOS балла?»

Нет. Один балл может быть получен. Однако должно быть ясно, на какой вопрос о популяции он отвечает. Например:

- Баллы популяции Естественный охват
- Баллы Общая опора сравнения
- Баллы, актуальные для рынка

- Результат языковой справедливости

могут иметь отдельные записи. Значение числа возникает из ясности его знаменателя.

ОБЩЕЕ СУЖДЕНИЕ ГЛАВЫ 44

GEO-1000 содержит в своём названии «1,000», но ценность этих 1,000 людей не определяется только числом. Она возникает из ответов на следующие вопросы: Кто они? Из какой популяции они были выбраны? Могут ли они действительно использовать продукт? Где они живут? Какие языки они используют? Сколько раз один и тот же человек был учтён? Видны ли маленькие страны? Сохраняют ли большие страны свой реальный вес в популяции? Где учёт учитывает людей, до которых продукт не может дотянуться? Кого исключила инфраструктура панели? Какой результат измеряет потенциальный доступ, а какой измеряет фактическое воздействие? У большой страны может быть большая численность населения, имеющего право на продукт, и, следовательно, она заслуживает значительного веса в основной панели. Но если продукт там недоступен, аудит не должен выдумывать пользователей. Разрыв в доступе должен раскрываться отдельно.

Если страна очень мала, основной вес населения должен оставаться низким. Тем не менее, он не обязательно должен быть полностью невидимым. Его можно наблюдать с помощью панели мониторинга. Если у языка небольшое количество носителей, он может не определять глобальное среднее значение. Однако если в этом языке существует серьёзное искажение представления, оно также должно быть заметно. Если человек знает три языка, он может создавать три разных опыта. Однако это не три разных человека. Если продукт ИИ предоставляет определённое поведение только на платном плане, оценка бесплатных пользователей не отражает этот план. Панель населения людей подходит для потребительского чат-продукта. Та же модель популяции может быть не подходящей для API или корпоративного продукта. Поэтому пятый закон измерений NOMOS выглядит следующим образом:

> *Даже самая чувствительная оценка дает правильный ответ на неправильный вопрос, если знаменатель неверен.*

Его шестой закон таков:

> *Вес населения не заменяет видимость страны и языка.*

Его седьмой закон таков:

> *Человек, который не имеет доступа к [чему-либо], не был неправильно представлен; однако он является частью глобальной границы доступа.*

Его восьмой закон таков:

> *Человек, имеющий несколько аккаунтов, языков или сессий, не увеличивает численность населения.*

Девятый закон звучит следующим образом:

> *Глобальное измерение не означает, что все измерены; это честная запись того, кто измерен, а кто нет.*

Порядок раздела 5 из NOMOS

> *Не давайте мне только численность населения стран. / Покажите, какие люди реально могут использовать продукт.*

Не заполняйте большую страну фальшивыми пользователями, если продукта там нет.

Не давайте одному человеку маленькую страну и не объявляйте ответ этого человека голосом страны.

Не дублируйте человека столько раз, сколько у него аккаунтов, языков и сессий.

Не путайте официальный язык с реальным использованием языка людьми.

Не делайте бесплатного пользователя представителем платного плана или мобильного пользователя представителем веб-интерфейса.

Не называйте панель волонтёров случайным населением мира.

Не удаляйте пользователя с особыми потребностями из числа участников только потому, что инструмент не работает.

Не исключайте из точности измерений население, до которого продукт не доходит, и не скрывайте разрыв в глобальном доступе.

Не применяйте одинаковую долю к вопросу о цене и к вопросу о идентичности, или к глобальному бренду и местной лицензии.

Сначала определите продукт. / Затем определите вопрос. / Затем определите человеческую вселенную. / Затем ограничьте доступ, язык, возраст и интерфейс. / Затем объясните, кого вы не можете выбрать. / И только после этого распределите 1,000 человек.

Заключительное предложение главы

В GEO-1000 справедливость заключается не в том, чтобы давать одинаковое количество людей каждой стране; она заключается в том, чтобы представлять каждого человека в правильной целевой популяции, с соответствующим весом, не скрывая места, где они не измеряются.

Нормативное ядро

Каждый аудит популяции GEO ДОЛЖЕН определить, до проверки результатов, ограниченную по времени и определённую для конкретного продукта целевую пользовательскую совокупность для каждого: - ИИ-продукта и пользовательского интерфейса, - страны или юрисдикции, - возраста и условий законного доступа, - условий аккаунта или плана, - языка и локали, - интерфейса, - семейства промптов, - и цели измерения. Общая численность постоянного населения НЕ ДОЛЖНА рассматриваться как совокупность пользователей, допущенных к использованию продукта, без обоснованных корректировок права на участие. Недоступность продукта ДОЛЖНА быть зафиксирована как ограничение охвата доступа, а не оценена без раскрытия как некорректное представление или тайно исключена из глобального охвата. Потенциальный доступ, активное использование, фактическое воздействие, Естественный охват, Общая область поддержки, актуальность на рынке и наблюдаемость в рамках исследования ДОЛЖНЫ оставаться различными понятиями целевой совокупности. Лица, аккаунты, сессии, языки и наблюдения за продуктом НЕ ДОЛЖНЫ рассматриваться как взаимозаменяемые единицы популяции. Страны с избыточной выборкой, языки и группы по признаку доступности ДОЛЖНЫ быть перевзвешены для оценки популяции, при этом оставаться отдельно видимыми для анализа справедливости и охвата. Наблюдения Country Observer МОГУТ устанавливать географический охват наблюдений, но НЕ ДОЛЖНЫ представляться как статистически надёжные оценки на уровне страны. Участников НЕ ДОЛЖНЫ обязывать обходить ограничения по продукту, географии, возрасту, идентичности или аккаунту, чтобы войти в официальную панель населения. Каждая рамка целевой совокупности ДОЛЖНА быть версионирована, опираться на источники, учитывать неопределённость и быть отнесена к ответственному лицу или организации.

ГЛАВА 06

Пропорциональная населению выборка GEO-1000

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K04; K05; K07; K08

Граница главы

Раздел 5 установил следующее положение:

> *Целевая пользовательская аудитория каждого продукта ИИ должна определяться отдельно. Общая численность населения страны не является численностью пользователей, имеющих право на продукт в этой стране.*

Теперь мы подходим к следующему вопросу:

> *Из определённой численности подходящих пользователей для GEO-1000, сколько человек будет выбрано, из какой страны, на каком языке и каким методом?*

Подходящая численность пользователей для ИИ-продукта составляет:

- 40% в стране A,
- 25% в стране B,
- 15% в стране C,
- 0.04% в маленькой стране

Если доступно, как следует распределять действительные наблюдения 1,000? Как дробные квоты будут преобразованы в целые числа? Будут ли страны с долей населения меньше одного человека полностью исключены? Будет ли нарушено взвешивание по населению, если каждому населению будет выделен хотя бы один человек? Можно ли объединить малые страны в отдельный пул? Если некоторые языки будут перепредставлены для языковой справедливости, как будет рассчитан глобальный балл? Если целью является 100 действительных наблюдений в стране, сколько людей следует пригласить? Кто заменит участника, если он не выполнит задание? Можно ли заменить пользователя, дающего неправильные ответы? Если один и тот же человек тестирует более одного продукта ИИ, будет ли выборка по-прежнему независимой?

Даже имея необработанную выборку из 1,000 человек, могут ли сильно неравные веса снизить эффективную статистическую информацию ниже уровня 1,000 независимых, равноосновных наблюдений? Это не просто организационные детали. Дизайн выборки определяет, насколько большое влияние каждый человек оказывает на показатель GEO. Дефектная выборка:

- может неверно взвешивать правильные ответы,
- может символически переоценивать малые страны,

- может делать большие популяции меньше, чем они есть на самом деле,
- может делать низкоресурсные языки невидимыми,
- может представлять добровольных пользователей технологий так, как если бы они были населением всего мира.

Его можно корректировать для достижения желаемого результата после просмотра результатов. Ранее установленная архитектура аудита требует, чтобы каждый аудит был измеримым, а у модели/продукта была указана дата, страна, язык, набор запросов, количество повторов и запись измерений. Этот раздел применяет ту же дисциплину к дизайну выборки. Поскольку подлинные доказательства требуют указания ограничений, контекста и времени, это также требует демонстрации того, как была создана выборка. Этот раздел:

- как будут распределены 1,000 действительные наблюдения основной панели населения,
- как будут установлены страты и ячейки,
- как будут управляться маленькие страны,
- как дроби будут преобразованы в целые числа,
- как будет учитываться переотбор,
- как разделить приглашение, резервные и действительные объекты наблюдения,
- как установить веса для проектирования и анализа,
- почему эффективный размер выборки может отличаться от исходного числа

определяет. В этой главе пока не рассматривается:

- из какого поставщика панели будут отобраны участники,
- подробности реализации полевого этапа при добровольной или вероятностной выборке,
- проверка и оценка ответов,
- метод окончательного доверительного интервала,
- все формулы оценивания NOMOS

не полностью окончательные. Задача раздела 6 заключается в следующем:

> *Преобразовать определенную целевую популяцию в зафиксированную, взвешиваемую и воспроизводимую выборку GEO-1000 до получения исхода.*

NOMOS Вызов

Вы создаете глобальную панель из 1,000 человек для продукта на основе ИИ. Доступные для продукта популяции десяти стран следующие:

Страна	Доступное количество пользователей
A	400,000,000
B	250,000,000
C	150,000,000
D	100,000,000
E	60,000,000
F	30,000,000
G	8,000,000
H	1,500,000
I	400,000

Страна	Доступное количество пользователей
J	100,000
Итого	1,000,000,000

Сырые квоты в соответствии с чистой пропорцией населения:

Страна	Сырая квота
A	400.0
B	250.0
C	150.0
D	100.0
E	60.0
F	30.0
G	8.0
H	1.5
I	0.4
J	0.1

Половину человека выбрать нельзя. 0.4 человек нельзя отвезти в страну. Если округлить сырые квоты до ближайшего целого числа, общая сумма может превысить или не достигнуть 1,000. Если дать хотя бы одного человека каждой стране, будут представлены страны I и J в большей степени по сравнению с их долей в населении.

Передача одного человека стране I увеличивает её теоретическую долю 2.5 раз. Передача одного человека стране J увеличивает её теоретическую долю 10 раз. Это экстремальное представление можно в значительной степени скорректировать. Но вы не можете получить показатель страны с одним человеком. Что произойдет, если вы не дадите ни одного человека странам I и J?

Если вы определяете эти страны как отдельные и обязательные страты, вероятность того, что некоторые люди из целевой популяции будут отобраны, становится нулевой. В этом случае утверждение о «вероятностной выборке, представляющей всю соответствующую мировую популяцию», слабеет. Теперь рассмотрим более опасное решение. Сначала вы собираете ответы.

Единственный пользователь в стране J дает неверный ответ. Затем вы говорите: «Размер выборки этой страны очень мал, давайте исключим ее из области исследования». В стране A много положительных ответов. Вы сохраняете ее с полным весом. Это уже не округление. Это изменение выборки, ориентированное на результат.

В другом случае вы увеличиваете процент недействительных записей в языковой группе с низкими результатами, используя резервных участников только до получения положительного ответа. Это также не является выборкой. Это механизм отбора ответов. Первое постановление этого раздела:

> *Выборка должна быть установлена до просмотра ответов III.*

Второе положение гласит:

> *Вероятность выбора участника не может зависеть от правильности его ответа.*

Её третье положение гласит:

> *Видимость малой страны и вес населения в глобальном рейтинге должны управляться отдельно.*

Четвертая статья гласит:

> *Пересэмплирование возможно; избыточное взвешивание не обязательно.*

Пятая его статья гласит:

> *Сырые записи 1,000 не всегда несут статистическую информацию, эквивалентную наблюдениям 1,000.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела заключается в преобразовании целевых действительных наблюдений 1,000 для каждого продукта ИИ и волны измерений в проверяемый дизайн выборки на основе пользовательной популяции, имеющей право на продукт. Раздел стандартизирует следующие различия:

- Целевая популяция против выбранной выборки
- Исходная квота против целочисленного распределения
- Основная популяционная панель против дополнительных панелей
- Видимость страны против веса страны
- Страты против экспериментальной ячейки
- Страты крупных стран против пула малых стран
- Пропорциональное распределение против непропорционального сверхвыборочного отбора
- Доля выборки против доли популяции
- Количество приглашений против цели по действительным наблюдениям
- Резервный участник против замены после результатов
- Недействительность сбора данных против неблагоприятного ИИ-результата
- Дизайнерский вес против веса результата
- Сырая выборка против эффективной выборки
- Выборка по странам против выборки по языкам
- Многоязычный пользователь против более чем одного человека
- Одностадийный отбор против многостадийного отбора
- Распределение Естественный охват против распределения Общая опора сравнения
- Моментальное распределение волн и распределение временных рядов
- Оценка населения и анализ по распознаванию стран или языков
- Основные наблюдения 1,000 с вспомогательными наблюдателями и панелями справедливости

В конце этого раздела каждый аудит должен быть в состоянии ответить на следующие вопросы:

> *Почему наблюдения 1,000 были распределены в таких количествах по стране, языку и группам пользователей?*
>
> *Как был определён вес каждого наблюдения в глобальном результате?*
>
> *Кто имел нулевую вероятность быть выбранным и почему?*
>
> *Сколько человек было приглашено, сколько завершили и сколько наблюдений осталось действительным?*

Были ли результаты выборки заморожены до просмотра?

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Основная выборка населения GEO-1000 должна основываться на стратифицированной схеме выборки, которая назначает известные или предсказуемые вероятности выбора людям в заранее замороженной рамке населения, подходящего для продукта, для каждого ИИ-продукта и волны измерений. Целевая величина основной панели населения по умолчанию:

По крайней мере 1,000 действительных наблюдений на каждый ИИ-продукт и основную волну измерений

сохраняется. Это число:

- не является количеством приглашенных людей,
- не является количеством аккаунтов,
- не является количеством необработанных скриншотов,

Это не сумма всех дополнительных панелей. 1,000 является целью для действительных наблюдений за продуктом, соответствующих протоколу в основной популяционной панели. Панели Country Observer, Language Fairness и Accessibility могут создавать дополнительные выборки. Эти дополнительные выборки:

- могут увеличивать общий объем полевых данных выше 1,000,
- могут использоваться с соответствующими весами в результатах глобальной популяции,

но не могут использоваться для скрытого уменьшения цели действительных наблюдений 1,000 основной панели.

3. БАЗОВОЕ ПРЕДСТАВЛЕНИЕ ВЫБОРКИ

Для конкретного продукта ИИ a, волны измерений w и целевой популяционной рамки, страты:

```
h∈ H_{a,w}
```

обозначены. Для каждой страты:

- Nh: оценка подходящей для продукта популяции в страте
- N: общая целевая популяция
- Wh: вес популяции страты
- n: целевая действительная основная панельная наблюдение
- qh: грубая пропорциональная квота
- nh: распределённые целые действительные наблюдения цели

является. Общая целевая популяция:

```
N= Σ_{h=1}^H N_h
```

Вес популяции страты:

```
W_h= N_h/ N
```

Грубая пропорциональная квота:

```
q_h= nW_h
```

является. Основное пропорциональное распределение при идеальных условиях:

```
n_h≈ nW_h
```

должно быть. Однако:

- qh может быть дробным,
- некоторые ячейки могут быть меньше одного человека,
- может потребоваться минимальная диагностическая выборка для подгрупп по языку и стране,
- также могут быть изучены небольшие группы высокого риска

Уровни заполнения полей могут различаться. Поэтому исходная квота не является прямым окончательным распределением.

4. ТРИ РАЗЛИЧНЫЕ ЦЕЛИ ДИЗАЙНА ВЫБОРОК

Та же выборка может не удовлетворять всем целям с одинаковой силой. Следует разделять три цели.

4.1. Глобальная оценка населения

Отвечает на вопрос:

> *Какой представительный результат увидит средний человек из глобальной популяции пользователей, подходящей для продукта?*

Приоритетным является вес населения.

4.2. Диагностика подгрупп

Отвечает на вопрос:

> *Какие искажения представления происходят в конкретных странах, языках, планах или группах пользователей?*

Может потребоваться сверхотбор малых групп.

4.3. Географическое и языковое покрытие

Отвечает на вопрос:

> *В каких странах и на каких языках наблюдается хотя бы реальное присутствие пользователей?*

Панели наблюдателей по странам и справедливости языка становятся важными. Эти три цели нельзя необъяснимо объединять в одной выборке.

5. ОСНОВНАЯ АРХИТЕКТУРА ВЫБОРКИ

Кандидат на основную архитектуру выборки для GEO-1000 состоит из четырех уровней.

5.1. Панель населения

Каждый продукт и волна ИИ имеют минимум 1,000 действительных наблюдений в качестве основной панели. Это позволяет получить оценку глобального населения.

5.2. Наложение наблюдателя по странам

Обеспечивает дополнительные географические наблюдения для небольших стран, которые не могут быть напрямую представлены в основной панели населения. Эти наблюдения могут указывать на:

- одиночном событии,
- статус доступа,
- потребности в будущих исследованиях

Они сами по себе не создают страновой рейтинг.

5.3. Дополнительная выборка справедливости языка

Генерирует дополнительные наблюдения для языков с ограниченными ресурсами или языков, недостаточно представленных в основной панели. Они перераспределяются согласно фактическим весам населения в глобальном результате.

5.4. Доступность - дополнительная выборка

Также измеряет пользователей вспомогательных технологий и доступности, которые недостаточно представлены в стандартной панели. Целевая вселенная и вес этих наблюдений фиксируются отдельно.

6. ЧТО ТАКОЕ СТРАТУМ?

Стратум — это сегмент населения, который управляется отдельно с точки зрения распределения выборки и вероятности отбора. Стратум может быть комбинацией одного или нескольких из следующих областей:

- Страна
- Регион
- Язык
- Локаль
- План ИИ
- Веб или мобильное
- Контролируемый или Панель пользователей в естественных условиях
- Статус доступности
- Тип пользователя
- Юрисдикция

Если каждое небольшое сочетание сделать отдельным слоем, выборка будет фрагментирована. Например: сочетание страна × язык × план × мобильный/веб × возрастная группа × тип пользователя может дать сотни или тысячи ячеек. Наблюдения 1,000 не предоставляют надежных оценок для большинства этих ячеек. Поэтому стратификация должна быть:

- существенной,
- измеримой,
- соответствующей численности выборки

должен быть.

7. СЛОЙ НЕ ТО ЖЕ САМОЕ, ЧТО ОТЧЕТНАЯ ЯЧЕЙКА

Он может отображаться отдельно на карточке группы результатов. Однако он не обязательно должен быть отдельным слоем в дизайне выборки. Например:

- небольшой язык,

- определенная группа с особыми возможностями,
- определенная возрастная группа

Его можно выбрать внутри более широкой страты в основной выборке. Позже его можно отдельно указать в анализе. Возможен и обратный вариант: два региона могут быть отдельно стратифицированы для повышения эффективности выборки. Однако в публичном отчете они могут быть объединены под одним региональным результатом. Дизайн стратификации и структура отчетности должны иметь отдельные записи.

8. ПОСЛЕДОВАТЕЛЬНОСТЬ СОЗДАНИЯ СТРАТЫ

Кандидатная последовательность для основной популяционной панели выглядит следующим образом:

- Продукт ИИ и пользовательский интерфейс
- Рамка популяции Естественный охват или Общая опора сравнения
- Страна или пул небольших стран
- Основной язык аудита или языковая группа
- План материалов/разделение интерфейса, если необходимо
- Рамка отбора участников

План и интерфейс не обязательно создавать в виде отдельных слоев для каждой страны. Это можно управлять через:

- подслой,
- квоту,
- случайное распределение

методы.

9. СЛОИ БОЛЬШИХ И МАЛЫХ СТРАН

Сырой квотой страны:

```
q_c= nN_c/ N
```

имеет такую форму. Если сырой квоты достаточно, страну можно сохранить непосредственно как отдельный слой. Это можно назвать:

слой страны, представляющей сама себя

Создание отдельных обязательных слоев для стран с очень маленькой сырой квотой может снизить эффективность выборки. Эти страны можно соответствующим образом:

- региональными,
- языковой семьи,
- режимом доступа к продукту,
- географической близостью

включать в пул малых стран.

10. ПУЛ МАЛЫХ СТРАН

Пул малых стран:

```
Пусть он будет представлен как B_r= {c_1, c_2, …, c_k}
```

Общая численность допустимого населения пула:

```
N_{B_r}= Σ_{c∈B_r} N_c
```

Основная квота выборки:

```
q_{B_r}= nN_{B_r}/ N
```

вычисляется как. В рамках пула выбор страны может осуществляться с вероятностью, пропорциональной размеру допустимого населения. Вероятность того, что страна c будет выбрана в пуле:

```
Pr(c | B_r) = N_c/ N_{B_r}
```

Возможно. Затем выбирается участник из выбранной страны. Этот подход:

- может предотвратить ситуацию, когда маленькие страны имеют нулевой шанс быть выбранными в целевой популяции,
- не гарантирует, что каждая страна будет выбрана в данной волне,

сохраняет глобальную оценку численности населения. Не выбранная страна в этой волне:

НЕ ВЫБРАНА В ВЫБОРКЕ НАСЕЛЕНИЯ

не означает: «Страна не является частью целевой популяции».

11. ОГРАНИЧЕНИЯ ОБЪЕДИНЕНИЯ МАЛЫХ СТРАН

Малые страны нельзя случайным образом объединять в один пул только из-за того, что их население небольшое. Страны в пуле должны быть схожи в тех аспектах, которые имеют значение:

- доступ к продукту,
- пользовательский интерфейс,
- структура языка или локали,
- юридический доступ,
- географический регион,
- метод сбора данных

должен соблюдаться. Страна без доступа к продукту не может быть размещена в том же основном пуле, что и страна с полным доступом. Различные условия использования не могут быть скрыты.

12. ПУЛ МАЛЫХ СТРАН НЕ СОЗДАЁТ ОТДЕЛЬНЫЙ СТРАНОВОЙ БАЛЛ

Один или два наблюдения, выбранные из пула, могут использоваться в глобальном вкладе пула в популяцию и могут указывать на отдельное событие в стране. Однако это не создаёт отдельный статистический балл для малой страны. Если требуется балл по стране, дополнительно:

- Наблюдатель страны,
- страно-специфическая выборка,
- многоэтапная кумулятивная выборка

может потребоваться.

13. РАСПРЕДЕЛЕНИЕ ПРОПОРЦИЙНО ЧИСТОМУ НАСЕЛЕНИЮ

Чистое пропорциональное распределение:
основано на формуле:

```
q_h= nW_h
```

Преимущества:

- это понятно для основной оценки населения,
- естественным образом придает вес большим группам населения,
- может уменьшать различия в весах,

может приближаться к самовзвешивающемуся дизайну. Ограничения:

- может отбирать очень мало малых стран и языков,
- может быть недостаточно для получения оценок подгрупп,
- может оставлять невидимыми группы высокого риска,

может создавать дробные и почти нулевые квоты. Следовательно, чистое пропорциональное распределение может быть сильным по умолчанию для основной Панели Населения. Оно не заменяет дополнительные панели.

14. ПРЕОБРАЗОВАНИЕ ДРОБЕЙ В ЦЕЛЫЕ ЧИСЛА

Сырые квоты часто бывают дробными. Общий объем распределения должен быть точно равен n. Метод наибольшего остатка может использоваться в качестве метода по умолчанию.

14.1. Первый шаг

Каждая сырая квота округляется вниз:

```
b_h= ⌊q_h⌋
```

14.2. Второй шаг

Количество нераспределенных наблюдений:

```
R= n− Σ_h b_h
```

вычисляется.

14.3. Третий шаг

Дробная часть каждой страты:

```
f_h= q_h− b_h
```

где f_h — дробная часть q_h.

14.4. Четвертый шаг

Оставшиеся R наблюдений добавляются по одному к слоям с наибольшими значениями fh.

14.5. Ситуация равенства

Если дробные части двух уровней равны:

- численность населения,
- предопределённый фиксированный порядок,
- зарегистрированное случайное начальное значение

Следует использовать независимое правило для разрешения равенства, независимо от результата. Страна не может быть выбрана на основании результатов.

15. АЛЛОКАЦИЯ СИНТЕТИЧЕСКИХ ЦЕЛЫХ ЧИСЕЛ

СИНТЕТИЧЕСКИЙ ОТРАЖЕНИЕ

Соответствующие популяции в начале раздела:

Страна	Подходящая популяция	Сырая квота
A	400,000,000	400.0
B	250,000,000	250.0
C	150,000,000	150.0
D	100,000,000	100.0
E	60,000,000	60.0
F	30,000,000	30.0
G	8,000,000	8.0
H	1,500,000	1.5
I	400,000	0.4
J	100,000	0.1

Если страны защищены как отдельные слои, округляйте вниз:

Страна	Округлённая вниз квота
A	400
B	250
C	150
D	100
E	60
F	30
G	8
H	1
I	0
J	0
Итого	999

Остается одно наблюдение. Оно добавляется в H, потому что наивысшая дробная часть равна 0.5 в стране H. Итоговое распределение:

Страна	Итоговая квота
A	400
B	250

Страна	Итоговая квота
C	150
D	100
E	60
F	30
G	8
H	2
I	0
J	0
Итого	1,000

Хотя эта конструкция определяет I и J как отдельные детерминированные слои, если она не даёт шанса быть выбранной, это может создать проблему с точки зрения полного вывода целевой вселенной. Более точное решение — небольшой сельский бассейн.

16. СИНТЕТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ С НЕБОЛЬШИМ СЕЛЬСКИМ БАССЕЙНОМ

Страны H, I и J:

```
B_small= {H, I, J}
```

Их нужно объединять в бассейне. Популяция бассейна:

```
1,500,000+400,000+100,000=2,000,000
```

Квота на пул:

```
1,000× 2,000,000/ 1,000,000,000= 2
```

становится. Основное распределение:

Стратум	Квота
A	400
B	250
C	150
D	100
E	60
F	30
G	8
Небольшой сельский бассейн	2
Итого	1,000

Веса отбора стран в пуле:

Страна	Доля населения в бассейне
H	75%
I	20%

Страна	Доля населения в бассейне
J	5%

По этим вероятностям выбираются два основных наблюдения популяции. В той же волне:

- оба наблюдения могут быть получены из H,
- один из H и один из I,
- редко из J

Это может происходить. Дизайн не делает J видимым в каждой волне, но это не снижает вероятность выбора J до нуля. Если необходимо наблюдать J, он добавляется отдельно в Панель Наблюдателей Страны. Наблюдение Страны не дает J веса один в 1,000 в основной глобальной выборке; J сохраняет свой истинный вес в основной панели.

17. МИНИМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ ЯЧЕЕК

Чисто пропорциональная квота может быть недостаточной для проведения диагностического анализа для некоторых подгрупп. Например, исходная квота для языковой ячейки может быть 8. Надежный языковой показатель нельзя получить при восьми наблюдениях. В этом случае есть три варианта:

- Не публиковать языковой показатель
- Объединить ячейку с похожими группами
- Избыточная выборка ячейки

Минимальное распределение ячеек должно быть определено до получения результатов.

18. ГРУППЫ ОТЧЁТОВ КАНДИДАТОВ

ГРУППЫ ОТЧЁТОВ КАНДИДАТОВ — ГРУППЫ ОТЧЁТОВ КАНДИДАТОВ

Приведённые ниже пороговые значения не являются окончательными научными решениями. Они должны быть откалиброваны через пилотное тестирование и внешнюю проверку.

1–29 Действительное наблюдение

Отчёт о единственном событии и открытии Страна или языковая ставка не публикуется

ТОЛЬКО НАБЛЮДЕНИЕ

30–99 Действительные наблюдения

Примерная оценка Увеличенная неопределённость Определённое суждение о соответствии не предоставляется

ЭКСПЕРИМЕНТАЛЬНО

100–199 Действительные наблюдения

Первая оценка подгруппы Неопределённость должна быть видна

ПРОВИЗОРНОЕ

200 и выше Действительные наблюдения

Более точная оценка подгруппы кандидат Влияние дизайна и веса оцениваются снова

СТАНДАРТНАЯ ОЦЕНКА КАНДИДАТА

Эти группы:

- уровень ошибок,
- ожидаемая точность,
- вариабельность веса,
- эффект проекта

не может автоматически считаться достаточной без учета.

19. ПЕРЕОТАБИРАНИЕ

У страны, языка или группы пользователей может быть больше наблюдений, чем доля их населения. Это называется переотабиранем. Его цели:

- повысить чувствительность подгруппы,
- сделать язык с низкими ресурсами заметным,
- изучить малую группу с высоким риском,
- расширить охват Country Observer,
- адекватно представить пользователей с ограниченными возможностями

возможны. Переотабирование не является ошибкой. Что является ошибкой:

> *Использовать долю выборки переотобранной группы, как если бы она была истинной долей населения.*

20. ДОЛЯ НАСЕЛЕНИЯ И ДОЛЯ ВЫБОРКИ

Доля населения для слоя h:

```
W_h= N_h/ N
```

Доля выборки:

```
S_h= n_h/ n
```

находится в форме. В случае чисто пропорционального распределения:

```
S_h≈ W_h
```

хорошо. При избыточной выборке:

```
S_h> W_h
```

Это возможно. При небольшой выборке:

```
S_h< W_h
```

Это возможно. Глобальная оценка населения должна сохранять значения Wh вместо Sh.

21. ОСНОВНАЯ РАСЧЕТНАЯ НАГРУЗКА

В простом стратифицированном отборе можно считать базовый вес каждого наблюдения в слое h следующим образом:

```
d_i= N_h/ n_h
```

Нормализованный вес слоя можно записать как:

```
a_h= W_h/ S_h
```

Нормализованный вес на уровне наблюдения может быть:

```
w_i= W_h/ n_h
```

Обеспечено, что сумма равна единице:

```
Σ_i w_i= 1
```

Взвешенная оценка бинарного перехода может быть рассчитана как:

```
θˆ= Σ_i w_iY_i
```

Здесь:

```
Yi=1: отклик, соответствующий критерию перехода

Yi=0: отклик, не соответствующий критерию
```

Это возможно. Финальная система NOMOS не обязательно должна использовать только бинарные результаты. Та же логика взвешивания может быть применена к многомерным оценкам откликов.

22. ПРИМЕР ИЗЛИШНЕЙ ВЫБОРОЧНОСТИ

СИНТЕТИЧЕСКИЙ ПРИМЕР

В подходящей популяции страны B:

- Язык X: 80 процентов
- Язык Y: 20 процентов

Так пусть будет. В панели популяции есть 250 наблюдений для страны B. Чистое распределение:

- Язык X: 200
- Язык Y: 50

Пусть 50 дополнительные наблюдения по справедливости языка будут собраны для Языка Y. Собранные наблюдения:

- Язык X: 200
- Язык Y: 100

Доля Языка Y в исходной выборке:

```
100/300 = 33.3%
```

Так оно и есть. Фактическая доля в населении составляет 20%. В учете мировой популяции Язык Y не может иметь вес 33.3%. Нормализованные веса внутри страт:

- Общий вес для Языка X: 80%
- Общий вес для Языка Y: 20%

сохраняется. Каждое из 100 наблюдений для Языка Y получает меньший индивидуальный вес, чем одно наблюдение для Языка X. В то же время все 100 наблюдений используются в отдельном диагностическом отчете для Языка Y. Этот подход: увеличивает видимость подгруппы, не искажает оценку мировой популяции.

23. РАСПРЕДЕЛЕНИЕ МНОГОЯЗЫЧНЫХ ПОЛЬЗОВАТЕЛЕЙ

Человек может подходить для более чем одной языковой ячейки. Один из предпочтительных подходов для основной панели населения заключается в назначении пользователю заранее определённого основного языка аудита. Основной язык аудита может определяться согласно следующим полям:

- Ежедневный язык использования ИИ
- Родной или домашний язык
- Рынок и локаль
- Компетенция в естественном использовании промпта
- Баланс панели

Если пользователь также будет тестироваться на других языках, эти наблюдения:

- составляют парную многоязычную панель,
- панель языковой справедливости

также обозначаются соответствующим образом. Одному и тому же человеку запрещено получать полное числовое представительство во всех языках.

24. РАСПРЕДЕЛЕНИЕ ЯЗЫКОВ В СТРАНЕ

Пусть подходящая популяция для языковых групп l в стране c равна: N_{c,l}. Если квота страны составляет: n_c, то исходная языковая квота:

```
q_{c,l}= n_c× N_{c,l}/ Σ_l N_{c,l}
```

можно рассчитать следующим образом. Однако, если языковые группы перекрываются:

```
Σ_l N_{c,l}> N_c
```

может быть возможным. В этом случае использование прямой доли создаёт двойной счёт. Варианты решения:

- классификация основного языка аудита,
- разделение человеческого веса,
- многоязычная общая модель,
- отдельная панель сопоставленных языков

может быть использована. Использованный метод следует зафиксировать.

25. РАСПРЕДЕЛЕНИЕ ПЛАНА И ИНТЕРФЕЙСА ВНУТРИ НАСЕЛЕНИЯ

Бесплатные и платные пользователи или веб- и мобильные платформы могут вести себя по-разному. Однако разделение каждой ячейки с учётом страны и языка отдельно на слои плана и интерфейса может чрезмерно фрагментировать выборку. Можно использовать три подхода.

25.1. Подраспределение пропорционально населению

Если известна фактическая распределение пользователей, распределение осуществляется согласно плану и долям интерфейса.

25.2. Сбалансированное экспериментальное распределение

Проба равного или большего объема проводится для сравнения различий между планом или интерфейсом. При глобальной оценке она перерассчитывается с учётом фактических долей использования.

25.3. Отдельные интерфейс продукта показатели

Бесплатные веб-продукты, платные веб-продукты и мобильные продукты отчётно рассматриваются как отдельные объекты аудита. Также можно получить объединённый результат. Метод, который будет использоваться, должен быть определён до проведения измерения.

26. НОМЕР ПРИГЛАШЕНИЯ И ЦЕЛЬ ДЕЙСТВИТЕЛЬНОГО НАБЛЮДЕНИЯ

Если действительной целью наблюдения для страты является nh, приглашение ровно nh человек обычно недостаточно. Потому что:

- некоторые люди не участвуют,
- некоторые не выполняют задание,
- некоторые записи остаются вне протокола,

некоторые пакеты доказательств становятся недействительными. Пусть ожидаемая норма завершения для стратума составляет rh, а ожидаемая норма действительности — vh. Целевая аудитория приглашения примерно:

```
m_h= [n_h/ (r_h v_h)]
```

можно рассчитать следующим образом. Пример:

- Цель по валидным наблюдениям: 100
- Ожидаемая норма завершения: 80 процентов
- Ожидаемая норма действительности: 90 процентов

```
m_h= [100/ (0.80× 0.90)] = 139
```

человек можно пригласить. Это не является гарантией. По мере продвижения полевых работ можно использовать заранее определённую резервную систему.

27. ПРЕДЕЛ НАДУВАНИЯ ПРИГЛАШЕНИЙ

Приглашение большего числа людей само по себе проблему смещения выборки не решает. Ранние откликавшиеся могут систематически отличаться от поздних. Только взятие первых 1,000 завершивших:

- часовой пояс,
- интерес к технологии,
- качество соединения,
- график работы

может вызвать предвзятость с точки зрения. Порядок приглашения и отбора:

- должен быть рандомизирован,
- должен быть зарегистрирован в рамках страты,

должен управляться на протяжении всего периода работы в поле.

28. СИСТЕМА АЛЬТЕРНАТИВНЫХ УЧАСТНИКОВ

Для каждой страты перед просмотром результатов должен быть создан упорядоченный резервный список. Резервный участник может быть активирован только по следующим причинам:

- Отсутствие ответа на приглашение
- Невозможность завершить задание
- Техническая или этическая неприемлемость
- Нарушение протокола требований
- Недействительность пакета доказательств
- Повторная регистрация того же лица

Резервный пользователь не может вмешиваться по следующей причине:

> *Предыдущий ответ ИИ пользователя дал неправильный, отрицательный или нежелательный результат.*

Неправильный ответ ИИ является допустимым системным результатом. Это не может быть причиной для замены участника.

29. СЛЕПОТА РЕЗУЛЬТАТОВ В РЕШЕНИИ О ДОПУСТИМОСТИ

Человек, принимающий решение о том, соответствует ли наблюдение протоколу максимально:

- должен оценивать ответ, не зная, правильный он или неправильный,
- или положителен он или отрицателен для бренда.

Сначала:

- Фиксация и проверка соответствия протоколу
- Затем должна быть проведена оценка по существу

в противном случае отрицательные ответы могут быть легче объявлены: «недопустимыми».

30. РЕЗЕРВНЫЙ ПОРЯДОК В СЛОЕ

Резервный участник должен быть, по возможности, из того же:

- страны,
- языка,
- интерфейс продукта,
- план,
- состояние панели

слой. Если резерва в том же слое не остаётся, применяется заранее определённая эскалация:

- Дополнительный случайный отбор в том же слое
- Заранее определённый пул верхнего слоя

- Оставление распределения незавершённым и предупреждение о масштабе
- Новая версия анализа

После получения результата привлечение пользователя из другой страны и закрытие разрыва не может быть выполнено тихо.

31. ЧТО ПРОИСХОДИТ, ЕСЛИ СЛОЙ СЛИШКОМ КОРОТКИЙ?

Он может опуститься ниже целевого показателя слоя. Пример:

- цель: 100
- допустимое наблюдение: 72

Правильные варианты:

- Расширить окно поля до заранее определённого размера
- Использовать резервный список
- Объединить ячейку с аналогичной и заранее определённой верхней ячейкой
- Сообщите результат с большой неопределенностью
- Не публикуйте результат ячейки
- Пересчитайте глобальный вес, используя подходящий метод

Неправильный вариант: полностью удалить слой с низкой оценкой из анализа.

32. ИЗМЕНЕНИЕ КВОТЫ ПОСЛЕ РЕЗУЛЬТАТА

Следующие изменения являются манипуляциями с образцами, ориентированными на результат:

- Снижение целевой квоты страны с низкой оценкой
- Увеличение квоты языка с высокой оценкой
- Аннулирование пользователей, предоставляющих отрицательные ответы
- Переопределение пула малых стран на основе результатов
- Исключение плана с критическими ошибками
- Добавление новых пользователей до достижения желаемого среднего

Каждое изменение квоты материала:

- новая версия распределения,
- обоснование,
- дата,
- лицо, принимающее решения
- Сохранение предыдущего результата

должно быть выполнено с помощью.

33. МНОГОЭТАПНАЯ ВЫБОРКА

Многоэтапная выборка может использоваться, когда прямой доступ ко всем индивидуумам невозможен. Пример:

- Выбор региона или пула стран
- Выбор страны
- Выбор панели или местного исследовательского партнера
- Выбор пользователя
- Назначение продукта AI или задачи

Общая вероятность выбора отдельного пользователя равна произведению условных вероятностей на каждом этапе:

```
π_i= π_r× π_{c|r}× π_{f|c}× π_{i|f}
```

Здесь:

- r: регион или небольшой пул стран
- c: страна
- f: рамка выборки или панель
- i: человек

Базовый проектный вес:

```
d_i= 1/π_i
```

Если вероятность выбора не может быть зафиксирована, проект не может быть представлен как полностью вероятностная выборка. [K04; K07]

34. РАСПРЕДЕЛЕНИЕ В ДОБРОВОЛЬНЫХ И НЕВЕРОЯТНОСТНЫХ ПАНЕЛЯХ

Настоящая вероятность-основанная мировая панель не всегда возможна. Можно использовать добровольную или профессиональную онлайн-панель. В этом случае квоты по странам и языкам все еще могут быть установлены в соответствии с населением, подходящим для продукта. Однако:

- Фактическая вероятность выбора участника может быть неизвестна,
- может происходить самовыбор,

весовые корректировки могут быть сделаны только на основе наблюдаемых переменных. Правильный статус:

Неверовятностная пользовательская панель, стратифицированная и откалиброванная по населению

может быть. Следующее утверждение использовать нельзя: «1,000 человек, выбранные полностью случайным образом из мировой популяции.»

35. АРХИТЕКТУРА ВЕСОВ

Конечный вес наблюдения может состоять из нескольких компонентов:

```
w_i= d_i× a_i^{NR}× g_i^{CAL}
```

Здесь:

- di: базовый вес дизайна
- aiNR: корректировка на неответ
- giCAL: коэффициент калибровки

35.1. Вес дизайна

Это обратное значение вероятности выбора участника в дизайне выборки.

35.2. Корректировка на неответ

Для выбранных лиц с похожими характеристиками:

- сколько ответили,
- сколько дали действительное наблюдение

может быть учтено. Для класса отклика выборки r:

```
a_r^{NR}= (Σ_{j∈S_r} d_j)/ (Σ_{i∈R_r} d_i)
```

Здесь:

- Sr: выбранная подходящая выборка
- Rr: те, кто ответил или предоставил действительное наблюдение

может быть. Коррекция на неполные ответы: полностью не устраняет необнаруженные искажения, должна быть объяснена вместе с её предположениями.

35.3. Калибровочный вес

Выборка к известным общенациональным суммам:

- страны,
- языка,
- возраст,
- план,
- пользовательский интерфейс.

Калибровка может применяться по этим измерениям, но только там, где существуют надежные внешние данные.

36. НОРМАЛИЗАЦИЯ ВЕСОВ

Веса:

- к общей численности населения,
- к общему размеру выборки,
- к индивидуальному

может быть нормализован. Используемая нормализация в формуле оценки должна быть объяснена. Индивидуально нормализованные веса:

```
w̃_i= w_i/ Σ_j w_j
```

получаются. Взвешенная оценка:

```
θ̂= Σ_i w̃_iY_i
```

вычисляется следующим образом.

37. ОГРАНИЧЕНИЕ ВЕСОВ

Некоторые участники могут получать очень высокие веса, потому что они представляют очень маленькие и недостаточно изученные группы. Эта ситуация:

- может увеличить оценочную дисперсию,

- влияние одного наблюдения,
- и волатильность показателей

может возрасти. Веса могут быть ограничены верхним пределом. Однако обрезка весов: может создать смещение в представлении населения, может вновь сделать маленькие группы невидимыми. Поэтому обрезка весов должна:

- быть определена до получения результатов,
- быть показана вместе с неограниченными результатами,

и сопровождаться анализом чувствительности.

38. ЭФФЕКТИВНЫЙ РАЗМЕР ВЫБОРКИ

Хотя сырое количество наблюдений равно n, если веса сильно различаются, фактическое количество информации может быть меньше. Эффективный размер выборки по Кишу:

```
n_eff= (Σ_i w_i)^2/ Σ_i w_i^2 [K08]
```

можно показать следующим образом. При равных весах:

```
neff≈n
```

получается. Если веса сильно дисбалансированы: neff<n. исследование GEO-1000:

- 1,000 действительных наблюдений,
- может иметь 620 эффективный размер выборки

. Для широкой аудитории: утверждение «Измерено 1,000 человек.» может выглядеть убедительно из-за точности. В основном отчете по меньшей мере:

- сырое количество действительных наблюдений,
- эффективный размер выборки,
- вариабельность весов

должны быть определены.

39. ЭФФЕКТ ДИЗАЙНА ВЕСА

Приблизительный эффект дизайна из-за весов:

```
DEFF_w≈ 1+ CV(w)^2
```

можно показать. Здесь:

```
CV(w) = sd(w) / mean(w)
```

Это коэффициент вариации весов и только приблизительное представление их эффекта. Общий эффект дизайна также может зависеть от:

- кластеризации,
- стратификации,
- повторных измерений,
- сопоставленных пользователей

Эти эффекты должны оцениваться отдельно.

40. КЛАСТЕРИЗАЦИЯ

То же самое:

- Из исследовательской панели,
- из страны,
- из языкового сообщества,
- с устройства или учреждения,
- от пользователя

Наблюдения могут быть похожи друг на друга. В этом случае 1,000 наблюдений не являются 1,000 полностью независимыми единицами информации. Кластеризация:

- в дизайне выборки,
- при расчете дисперсии,
- в эффективном размере выборки

должна учитываться.

41. ТЕСТИРОВАНИЕ НЕСКОЛЬКИХ ИИ-продуктОВ ОДНИМ ПОЛЬЗОВАТЕЛЕМ

Позволение пользователю тестировать несколько ИИ-продуктов позволяет измерить различия между моделями на одном и том же человеке. Преимущество: характеристики пользователя могут контролироваться, парное сравнение может быть усилено. Ограничение:

- наблюдения по продукту не являются независимыми,
- порядок выполнения задач может влиять на другие ответы,
- пользователь может перенести ответ, увиденный в первой системе, в следующую систему,

пользователи, которые могут получить доступ ко всем продуктам, могут отличаться от общей популяции. Следовательно, продукты:

- должны быть назначены в случайном порядке,
- предыдущий обмен ответами должен быть предотвращён,
- парный дизайн должен быть зафиксирован,

Анализ должен защищать зависимость. Каждый AI продукт снова должен иметь свою собственную цель 1,000 действительных наблюдений продукта. Если одни и те же 1,000 люди тестируют десять продуктов:

- 1,000 уникальных людей,
- 10,000 наблюдений продукта

можно сформировать. Эти два числа не одинаковы.

42. СБАЛАНСИРОВАННЫЙ ДИЗАЙН С ПРОПУЩЕННЫМИ БЛОКАМИ

Не всем пользователям нужно получать доступ ко всем AI продуктам. Например, если есть десять AI продуктов, каждому участнику можно назначить три продукта. Если назначения будут сделаны сбалансированным образом:

- каждому продукту 1,000 наблюдений,
- каждой паре продуктов достаточное количество совместных пользователей,
- широкое участие населения

можно достичь. В качестве кандидата для этого подхода:

Дизайн сбалансированных неполных блоков

может быть использован. Его преимущества:

- снижает нагрузку на участников,
- уменьшает риск выбора только активных пользователей ИИ,

обеспечивает сопоставимое сравнение между моделями. Точный дизайн будет подробно описан в Разделах 8 и 9.

43. Естественный охват РАСПРЕДЕЛЕНИЕ

Каждому продукту ИИ выделяется отдельное распределение в соответствии с его собственной Естественный охват популяцией. Следовательно, одна и та же страна:

- 200 для одного продукта,
- 0 для другого продукта,
- 50 для третьего продукта

можно наблюдать. Эта разница не является несправедливостью. Дело в том, что реальные доступные аудитории продуктов различаются. Поскольку оценки Естественный охват принадлежат различным популяциям, их не следует использовать отдельно для прямого ранжирования.

44. Общая опора сравнения РАСПРЕДЕЛЕНИЕ

Создается общая доступная аудитория для сравнения продуктов. Общая популяция:

```
U_CS= ⋂_{a∈A} U_a
```

можно рассматривать таким образом. Каждый продукт измеряется с использованием одинаковых весов популяции Общая опора сравнения. Этот подход:

- обеспечивает справедливое сравнение продуктов,
- одинаковый минимальный уровень по странам и языкам

гарантируется. Однако он не показывает естественные глобальные различия в доступе к продуктам. Поэтому результаты Естественный охват и Общая опора сравнения должны сообщаться вместе.

45. УПРАВЛЕНИЕ ОБРАЗЦОМ В РАЗЛИЧНЫХ ВОЛНАХ

Могут быть две различные цели в разных волнах измерений.

45.1. Результат текущего среза

В каждой волне используется текущая рамка населения, подходящего для продукта. Этот показатель отвечает на вопрос: «Что происходит сейчас в реальной глобальной вселенной доступа?»

45.2. Результат тренда с фиксированной рамкой

Весовые коэффициенты населения из первоначальной волны сохраняются. Этот показатель отвечает на вопрос: «Как изменилось представление в той же эталонной популяции с течением времени?»

45.3. Разделение двух результатов

Когда продукт расширяется на новые страны, текущий показатель может снизиться. Это снижение:

- обусловлено ухудшением продукта,
- из-за включения новых и труднодоступных популяций

Это может быть объяснено. Тренд с фиксированной рамкой может одновременно показывать улучшение. Следовательно:

- Снимок текущей рамки
- Тренд с фиксированной рамкой

Они являются отдельными результатами.

46. ОБНОВЛЕНИЕ ПАНЕЛИ И ТРАНСФОРМАЦИЯ

Среди волн:

- те же пользователи,
- совершенно новые пользователи,
- частично обновленная панель

пригодна к использованию.

46.1. Полная панель секции

Новые пользователи выбираются в каждой волне. Это может быть сильным для оценки популяции. Не измеряет индивидуальные изменения.

46.2. Продольная панель

Те же пользователи измеряются снова. Показывает индивидуальные изменения пользователей и продуктов. Панель может создавать эффекты усталости и обучения.

46.3. Ротационная панель

Часть участников сохраняется, а часть обновляется. Обеспечивает как тенденции, так и представление новой популяции. Используемая структура панели должна быть видна на карточке результата.

47. ФИКСАЦИЯ ВЫБОРКИ

Следующие записи должны быть зафиксированы перед началом каждой волны:

- AI продукт
- пользовательский интерфейс
- Версия целевой популяции
- Естественный охват страны

- Общая опора сравнения страны
- Страты
- Небольшие страны-пулы
- Исходные квоты
- Целочисленные распределения
- Распределение по языкам
- План сверхвыборки
- Цели приглашений
- Резервные списки
- Метод выбора
- Формула весов
- Правило слепоты результатов
- Эскалация нехватки
- План анализа
- Метод декодирования
- Случайное зерно, если используется
- Ответственное одобрение человеком

Для этого файла:

NOMOS Фиксация распределения выборки

может быть дано имя.

48. ИЗМЕНЕНИЯ ДОПУСКАЮТСЯ ПОСЛЕ ЗАПЕРАНИЯ ОБРАЗЦА

Следующие события могут потребовать новую версию:

- Прекращение доступа к продукту в стране
- Серьезная проблема безопасности на практике
- Существенная ошибка в данных о населении
- Поставщик панели не может доставить в определенную страну
- Изменение требований к исследовательской этике
- Критическая ошибка в измерительном инструменте

Изменение:

- должно быть обосновано,
- датированная,
- утверждено,
- выполняться с сохранением старой версии.

Сила или слабость результатов не является основанием для изменения.

49. СТАТУСЫ КАЧЕСТВА ОБРАЗЦА

Волна образца может иметь один из следующих статусов:

SQ-0— НЕ ПРЕДНАЗНАЧЕН

Документированный образец дизайна отсутствует.

SQ-1— НАБЛЮДЕНИЕ ДЛЯ УДОБСТВА

Проводится наблюдение для удобства или добровольное наблюдение. Обобщение на население невозможно.

SQ-2— КВОТНО-КОНТРОЛИРУЕМОЕ

Существуют квоты по странам и языкам. Вероятности выбора полностью неизвестны.

SQ-3— ВЗВЕШЕННАЯ ПАНЕЛЬ

Существует стратифицированная панель и калибровка по населению. Ограничения по охвату и самовыбору сохраняются.

SQ-4— ОСНОВАННОЕ НА ВЕРОЯТНОСТИ ИЛИ ГИБРИДНОЕ

Известны вероятности выбора или существует надёжный гибридный дизайн.

SQ-5— РЕПЛИКИРОВАННОЕ И КАЛИБРОВАННОЕ

Дизайн был реплицирован в разных волнах и внешних приложениях; искажения и поведение дисперсии были откалиброваны. Эти статусы могут повлиять на класс доверия публичного рейтинга.

50. СИНТЕТИЧЕСКОЕ КОНЕЧНОЕ РАСПРЕДЕЛЕНИЕ GEO-1000

СИНТЕТИЧЕСКАЯ ДЕМОНСТРАЦИЯ МЕТОДОЛОГИИ / Следующие страны, население, языки, продукт ИИ и расчеты полностью вымышлены. Для продукта Synthetic Orion AI, пусть подходящее население для Естественный охват составляет 1 миллиардов человек. Цель Главной панели населения:

```
n=1,000
```

50.1. Распределение по странам и малым странам

Стратум	Подходящая популяция	Доля населения	Основная квота
A	400,000,000	40%	400
B	250,000,000	25%	250
C	150,000,000	15%	150
D	100,000,000	10%	100
E	60,000,000	6%	60
F	30,000,000	3%	30
G	8,000,000	0.8%	8
Небольшой сельский бассейн	2,000,000	0.2%	2
Итого	1,000,000,000	100%	1,000

50.2. Распределение языков в Стране B

Допустимое население в Стране B:

- Язык X: 80 процентов
- Язык Y: 20 процентов

Распределение по Панели населения:

- Язык X: 200
- Язык Y: 50

Дополнительное распределение по принципу языковой справедливости: Язык Y: 50 Целевой показатель в Стране B:

- Язык X: 200
- Язык Y: 100
- Итого: 300

Однако глобальные весовые коэффициенты населения Панели населения:

- Язык X: 80% страны B
- Язык Y: 20% страны B

остаются такими.

50.3. Приложение наблюдателя страны

Существуют два вероятностных наблюдения для H, I и J в малом пуле стран в Панели Населения. Панель наблюдателей страны также:

H: 1

I: 1

J: 1

может нацеливать наблюдения. Эти три наблюдения:

- увеличивают географическое охватывание,
- показывают события отдельных стран,

не создают оценку страны.

50.4. Дополнение по доступности

Если установлено, что основная панель недостаточно охватывает пользователей с ограниченными возможностями, можно провести дополнительные наблюдения 40. Эти пользователи: калибруются в соответствии с соответствующими группами населения, дают отдельный результат Панели по доступности.

50.5. Общий объем поля

Панель	Текущая цель
Панель населения	1,000
Дополнение по языковой справедливости	50
Дополнение по наблюдателям страны	3
Дополнение по доступности	40
Общая валидная цель поля	1,093

Имя GEO-1000 по-прежнему действителен. Потому что:

1,000 является действительной целевой наблюдением для основной панели населения.

Дополнительные панели справедливости и охвата могут быть добавлены к этому.

50.6. Объем приглашений

На Панели населения в среднем:

- завершение: 82 процентов
- действительность: 92 процентов

пусть это будет ожидаемо. Требуется приглашение:

```
[1,000/ (0.82× 0.92)] = 1,326
```

может быть. С дополнительными панелями общее количество приглашений может превысить 1,500. Поэтому:

GEO-1000— это не только метод отправки приглашений 1,000 людям.

50.7. Результат поля

На панели населения:

- 1,326 человек приглашено,
- 1,080 задач выполнено,
- 1,012 записей признано действительными с точки зрения протокола,
- первые 1,000 были включены в основной анализ согласно заранее определенному случайному/резервному порядку

Хорошо. Оставшиеся 12 действительных записей:

- резервная проверка,
- анализ чувствительности,
- повторная проверка

Это может быть сохранено. Нельзя определить, какой 1,000 будет выбран, глядя на результаты.

51. НОРМАТИВНЫЙ АЛГОРИТМ РАСПРЕДЕЛЕНИЯ

Последовательность операций по умолчанию для основной панели GEO-1000 выглядит следующим образом:

Шаг 1— Зафиксировать продукт и идентификатор волны

Определить, какой продукт ИИ, интерфейс и дата измеряются.

Шаг 2— Зафиксировать версию целевой популяции

Используется рамка популяции, допустимой для продукта, из раздела 5.

Шаг 3— Определить основную текущую целевую наблюдаемую величину

Целевой кандидат по умолчанию:

```
n=1,000
```

Шаг 4— Определить материальные слои

Определяются страны, язык и, при необходимости, слои плана/интерфейса.

Шаг 5— Настроить пулы для малых стран

Страны, которые будут получать почти нулевые квоты, включаются в соответствующие пулы.

Шаг 6— Рассчитать исходные пропорциональные квоты

```
qh=nWh
```

Шаг 7— Выполнить целочисленное распределение

Применяется метод наибольшего остатка или эквивалентный заранее объявленный метод.

Шаг 8— Добавить план сверхвыборки

Дополнения для наблюдателя страны, языковой справедливости и доступности сохраняются отдельно.

Шаг 9— Рассчитать количество приглашений и резервных номеров

Используются ожидаемые показатели завершения и действительности.

Шаг 10— Рандомизировать порядок выбора

Основные и резервные списки фиксируются.

Шаг 11— Сохранить формулу весов

Определяются проектные, ответные и калибровочные веса.

Шаг 12— Подтвердить блокировку выборки

Запись версии и целостности создается перед просмотром ответов ИИ.

52. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH06-N01

Распределение образцов GEO-1000 должно быть создано и версионировано перед просмотром ответов ИИ.

CH06-N02

Основной целевой показатель Панели Населения должен быть определен как не менее 1,000 действительных наблюдений для каждого продукта ИИ и основной волны измерений; если используется другой показатель, необходимо объяснить обоснование.

CH06-N03

Приглашённое лицо, лицо, выполняющее задачу, и действительные номера наблюдений должны быть различимы друг от друга.

CH06-N04

Основное распределение должно основываться на населении, соответствующем продукту, определённом в Разделе 5.

CH06-N05

Квоты по странам или языкам должны рассчитываться на основе доли соответствующего населения, имеющего право на продукт, а не общей численности населения.

CH06-N06

Метод преобразования дробных квот в целые числа должен быть определен заранее.

CH06-N07

Разрешение неоднозначностей при округлении должно быть независимым от результатов.

CH06-N08

Небольшие страны с положительной целевой популяцией не следует оставлять в качестве отдельных детерминированных страт с нулевой вероятностью; следует использовать соответствующий метод объединения или вероятностного отбора.

CH06-N09

Небольшие государственные пулы должны быть обоснованы с точки зрения доступа к продукту и финансовых характеристик населения.

CH06-N10

Наблюдения странового наблюдателя также должны быть зафиксированы без изменения основного веса населения.

CH06-N11

Наблюдение странового наблюдателя не может быть представлено в виде оценки страны.

CH06-N12

Страны, языки или группы пользователей с избыточной выборкой должны быть возвращены к истинным весам целевой популяции при глобальной оценке.

CH06-N13

Выборочная пропорция не может использоваться так же, как и пропорция в популяции.

CH06-N14

Один и тот же человек не может быть дублирован в глобальном человеческом населении из-за множества языков, учетных записей, сессий или продуктов.

CH06-N15

Недопустимые наблюдения следует классифицировать, не основывая причину недопустимости на содержании результатов.

CH06-N16

Некорректные, отрицательные, отвергающие или критические ответы ИИ не могут быть причиной замены участника.

CH06-N17

Резервные списки участников должны быть созданы внутри страты до того, как будут видны результаты.

CH06-N18

Разрывы между уровнями не могут быть тихо закрыты другой страной или языком.

CH06-N19

Изменения квоты после результатов, уровня, пула маленькой страны или веса требуют новой версии анализа.

CH06-N20

Все компоненты весов и метод нормализации должны быть опубликованы.

CH06-N21

Предположения корректировки на неотклик должны быть объяснены.

CH06-N22

Если используется усечение веса, необходимо сохранить неусечённый результат и анализ чувствительности.

CH06-N23

В основном отчёте следует показывать соответствующий эффективный размер выборки вместе с количеством наблюдений.

CH06-N24

При расчёте дисперсии следует учитывать кластеризацию, повторение и зависимости между парными пользователями.

CH06-N25

Дизайн, при котором один и тот же человек тестирует несколько продуктов ИИ, не может быть представлен как полностью независимая выборка пользователей.

CH06-N26

Каждый продукт ИИ должен отдельно достигнуть своей цели в 1,000 действительных наблюдений продукта.

CH06-N27

Распределения Естественный охват и Общая опора сравнения должны содержать отдельные записи выборки.

CH06-N28

Текущие и фиксированные результаты по волнам не могут быть объединены без пояснения.

CH06-N29

Если выборка недостаточна для подгруппы, правильный статус должен быть НЕДОСТАТОЧНАЯ ВЫБОРКА, ЭКСПЕРИМЕНТАЛЬНО либо НЕ ОЦЕНЕНО.

CH06-N30

Размер исходной выборки не может использоваться как окончательный показатель точности, игнорируя веса и эффект дизайна.

CH06-N31

Весовые коэффициенты дизайна не могут применяться в добровольной или непроизвольной панели так, как если бы известны были вероятности отбора.

CH06-N32

Выборка не может быть изменена постфактум для исключения групп с низкими показателями или для увеличения групп с высокими показателями.

CH06-N33

Наблюдения из основной панели населения и дополнительных панелей должны быть отдельно обозначены по назначению и весу.

CH06-N34

Фиксация выборки должна иметь ответственного человека или институционального владельца.

53. ФОРМЫ ОШИБОК

CH06-F01— УСТАНОВИТЬ КВОТУ ПОСЛЕ РЕЗУЛЬТАТА

Количество стран и языков определяется после просмотра ответов ИИ.

CH06-F02— ИСПОЛЬЗОВАТЬ РЕЗЕРВНУЮ ВЕРСИЮ ДО ЛУЧШЕГО ОТВЕТА

Отрицательные ответы заменяются новым пользователем, даже если они действительны.

CH06-F03— СЧИТАТЬ 1,000 ПРИГЛАШЕНИЯ КАК 1,000 ДЕЙСТВИТЕЛЬНЫЕ НАБЛЮДЕНИЯ

Недозавершенность и недействительность скрыты.

CH06-F04— СБРОСИТЬ СТРАНУ ДО НУЛЯ С ЧИСТОЙ ОКРУГЛЕНИЕМ

Вероятность выбора страны с положительной целевой популяцией остается равной нулю.

CH06-F05— ДАТЬ ОДНОМУ ЧЕЛОВЕКУ В КАЖДУЮ СТРАНУ И НАЗВАТЬ ЭТО ПРОПОРЦИЕЙ НАСЕЛЕНИЯ

Несоразмерность, вызванная минимальной квотой, скрыта.

CH06-F06— УЧИТЫВАЯ НАБЛЮДЕНИЯ СТРАНЫ-КОРРЕСПОНДЕНТА КАК НАБЛЮДЕНИЯ ЗА НАСЕЛЕНИЕМ

Дополнительное наблюдение за небольшой страной получает полный вес в основной панели людей 1,000.

CH06-F07— СОЗДАНИЕ ПУЛА МАЛЫХ СТРАН НА ОСНОВЕ РЕЗУЛЬТАТОВ

Страна с низкими показателями переносится в другой пул или исключается из охвата.

CH06-F08— ОБЪЕДИНЕНИЕ НЕСОВМЕСТИМЫХ СТРАН В ОДИН ПУЛ

Страны, которые различаются по доступу к продукту, языку или юрисдикции, объединяются исключительно из-за того, что они маленькие.

CH06-F09— ПРЕУВЕЛИЧЕНИЕ ВЕСА ЧРЕЗМЕРНЫХ ВЫБОРОК

Определяет глобальный результат с долей выборки выше, чем доля населения по языку или стране.

CH06-F10— УМНОЖЕНИЕ МНОГОЯЗЫЧНЫХ ЛИЦ

Человек считается отдельной единицей населения в каждой языковой ячейке.

CH06-F11— РАССМОТРЕНИЕ ДОЛИ ВЫБОРКИ КАК ДОЛИ НАСЕЛЕНИЯ

Сырые числа ответов прямым образом преобразуются в глобальный балл.

CH06-F12— УПОРЯДОЧИВАНИЕ ПРИГЛАШЕНИЙ ПО ПОРЯДКУ УДОБСТВА

Пользователи, которые отвечают первыми и быстро определяют панель.

CH06-F13— ИСПОЛЬЗОВАНИЕ РЕЗЕРВОВ STRATUM-OUT

Отсутствующие наблюдения по стране или языку тихо дополняются участниками из другой группы.

CH06-F14— СВЯЗЬ НЕПРАВИЛЬНОСТИ С СОДЕРЖАНИЕМ ОТВЕТА

Некорректный ответ ИИ технически объявляется недействительным.

CH06-F15— ЗАМЕНА ОТКАЗОВ НОВЫМ ПОЛЬЗОВАТЕЛЕМ

Отказ системы удаляется из знаменателя, даже если это фактический результат пользователя.

CH06-F16— СКРЫТИЕ ФОРМУЛЫ ВЕСА

Общественный рейтинг не показывает, с какой популяцией и весами выборки он был получен.

CH06-F17— ОТРЕЗАНИЕ ВЕСОВ ДЛЯ УДАЛЕНИЯ МАЛЫХ ГРУПП

Малые группы с высокими весами обрезаются, потому что результат выглядит плохо.

CH06-F18— СКРЫТИЕ ЭФФЕКТИВНОЙ ВЫБОРКИ

Даже если исходные записи 1,000 уменьшаются до 500–600 точек данных из-за весов, это представляется как «точность на 1,000 человек».

CH06-F19— РАСЧИТЫВАНИЕ КЛАСТЕРИЗАЦИИ КАК НЕЗАВИСИМОСТИ

Наблюдения от одной панели, учреждения или пользователя считаются полностью независимыми.

CH06-F20— РАСЧИТЫВАНИЕ ОДНИХ И ТЕХ ЖЕ 1,000 ЛЮДЕЙ КАК ДЕСЯТИ ТЫСЯЧАМ НЕЗАВИСИМЫХ ЛЮДЕЙ

Десять наблюдений за продуктом ИИ описываются так, как если бы популяция была в десять раз больше.

CH06-F21— РАСЧИТЫВАНИЕ Общая опора сравнения КАК Естественный охват

Обычная выборка для сравнения представлена так, как если бы это была фактическая глобальная аудитория продукта.

CH06-F22— ПРЯМОЕ РАНЖИРОВАНИЕ Естественный охват РЕЗУЛЬТАТОВ

Результаты продукта в различных целевых аудиториях сравниваются так, как если бы они были из одной и той же вселенной.

CH06-F23— СКРЫТИЕ ИЗМЕНЕНИЙ В АУДИТОРИИ В ЦИКЛАХ

Несмотря на то, что новые страны или группы пользователей включаются в результаты, тренд интерпретируется только как изменение качества продукта.

CH06-F24— ПОЛУЧЕНИЕ ОПРЕДЕЛЕННЫХ ОЦЕНОК ИЗ НЕДОСТАТОЧНЫХ ВЫБОРК ДЛЯ ПОДГРУППЫ

Несколько десятков наблюдений публикуются как определенные оценки для страны или языка.

CH06-F25— БЕСШУМНОЕ ИЗМЕНЕНИЕ ЗАФИКСИРОВАННОЙ ВЫБОРКИ

Определения квот, веса или пула обновляются без записи версии.

CH06-F26— СЛЕПОЕ УМНОЖЕНИЕ КОЭФФИЦИЕНТОВ ОТБОРУ

Расчеты по правомочной популяции и квотам игнорируют характеристики зависимой популяции.

CH06-F27— ЗАМЕНА ОСНОВНОЙ ПАНЕЛИ ДОПОЛНИТЕЛЬНОЙ ПАНЕЛЬЮ

Наблюдения за языковой справедливостью или наблюдения Страны используются для уменьшения основной целевой популяции 1,000.

CH06-F28— УДАЛЕНИЕ ЯЧЕЕК С НИЗКИМ БАЛЛОМ, КОГДА ПАНЕЛЬ МАЛА

Неполные образцы исключаются, поскольку они неблагоприятны с точки зрения результатов.

CH06-F29— ИСПОЛЬЗОВАНИЕ ВЕСОВ КАК НАУЧНОЙ ИЗЫСКАННОСТИ

Утверждается, что проблема представления решается только добавлением веса без объяснения смещения при отборе и охвате.

CH06-F30— ПОЛАГАНИЕ, ЧТО GEO-1000 ДОЛЖЕН БЫТЬ СПРОШЕН ТОЧНО У 1,000 ЛЮДЕЙ

Игнорируются приглашения, дополнительные панели, недействительность, повторение и структура эффективной выборки.

54. ПРОЦЕДУРА АУДИТА

Шаг 1— Записать Цель Выборки

Глобальная оценка численности населения делится на диагноз подгруппы, охват стран, языковую справедливость, доступность и цели сравнения продукта.

Шаг 2— Проверить Версию Целевой Популяции

Файл населения и запись даты в разделе 5 проверяются.

Шаг 3— Определить основную текущую целевую наблюдаемую величину

Записывается целевой показатель по умолчанию GEO-1000 или обоснованная альтернатива.

Шаг 4— Установить Страты

Определяются различия в стране, языке и материале пользовательский интерфейс.

Шаг 5— Выделить Бассейны Малых Стран

Для каждого бассейна:

- стран,
- население,
- обоснование сходства,
- внутренний метод отбора

записываются.

Шаг 6— Рассчитать Исходные Квоты

```
qh=nWh
```

Используется формула или обоснованная альтернатива.

Шаг 7— Применить целочисленное назначение

Правила округления и разрешения ничьих зарегистрированы.

Шаг 8— Добавить дополнительную сверхвыборку

Цели Наблюдателя страны, Языковой справедливости и Доступности определяются отдельно.

Шаг 9— Рассчитать объем приглашений и резерва

Используются ожидания завершения и действительности.

Шаг 10— Заморозить выбор и порядок резерва

Участники выбираются или упорядочиваются до того, как результаты станут известны.

Шаг 11— Рассчитать проектные веса

Вероятности выбора и веса слоев зарегистрированы.

Шаг 12— Определить процедуру для молчания и неотвечания

Ранее объявленный ответ записывается в разрыв слоя.

Шаг 13— Создайте образец замка

Добавляются идентификатор документа, версия, дата и запись целостности.

Шаг 14— Пересмотреть процесс валидности после полевых действий

Проверено, что решения о достоверности даются независимо от точности ответов.

Шаг 15— Генерируйте окончательные веса

Конструкция разделена на компоненты отклика и калибровки.

Шаг 16— Рассчитать эффективный размер выборки

Разница между исходным числом и объёмом информации сообщается.

Шаг 17— Сравните распределение и фактические данные

Планируемые и фактические показатели наблюдений отображаются постепенно по клетке.

Шаг 18— Пересмотр изменений после результатов

Проверяются изменения квот, пула, веса и объема.

55. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Идентичность ИИ-продукта; волна измерений; версия целевой популяции; цель Популяционной панели; цель Наблюдателя страны; цель Языковой справедливости; цель Панели доступности; список страт; популяции страт; веса популяции; необработанные квоты; метод округления; записи дробного распределения и остатка; правило разрешения равенства; пулы малых стран; вероятности выбора внутри пула; распределение по странам и языкам; распределение по плану и интерфейсу; обоснования сверхвыборки; цели приглашений; предположения о завершении; предположения о валидности; основные и резервные списки участников; запись выбора или рандомизации; записи неполного ответа; причины недействительных наблюдений; даты решения о валидности; проектные веса; корректировки на отсутствие ответа; калибровочные коэффициенты; политика обрезки весов; исходные и эффективные размеры выборки; кластеризация и записи сопоставленных пользователей; распределение Естественный охват; и распределение Общая опора сравнения.

Фиксированные и текущие веса кадров Сравнение планового и фактического распределения Фиксация образца Журнал изменений Ответственное лицо или учреждение

56. ЧЕК-ЛИСТ АУДИТА

Основная цель наблюдения открыта? Является ли цель 1,000 пригласительной или действительным наблюдением? Была ли выборка заморожена до того, как были получены ответы ИИ? Какая рамка популяции использовалась? На каком основании были установлены страты? Показаны ли необработанные квоты по странам и языкам? Понятен ли метод преобразования долей в целые числа? Было ли правило прерывания определено заранее? Равна ли вероятность выбора для малых стран нулю? Логически ли согласованы пула малых стран? Были ли наблюдения Country Observer включены в основным вес популяции? Был ли составлен оценочный показатель страны на основе наблюдения Country Observer? Если группы по языку и доступности были перепредставлены, были ли они повторно взвешены? Учтена ли одна и та же персона более чем как одна единица популяции? Приведены ли отдельно количество приглашений, завершивших и действительных наблюдений?

Были ли заранее созданы резервные списки? Был ли участник заменён из-за неправильного или отрицательного ответа ИИ? Было ли принято решение о достоверности до оценки точности материала? Был ли разрыв вкладок тихо закрыт из другой страны? Изменилась ли квота или пул после получения результатов? Можно ли воспроизвести весовые коэффициенты дизайна? Явны ли предположения о настройках отсутствующих ответов? Было ли проведено усечение весов и показан ли его эффект? Предоставляются ли вместе исходные и эффективные размеры выборки? Тестировал ли один и тот же пользователь более одного продукта ИИ? Сохранена ли эта зависимость в анализе? Достиг ли каждый продукт ИИ отдельно продукта-наблюдений 1,000? Были ли образцы Естественный охват и Общая опора сравнения разделены? Изменилась ли рамка популяции между волнами?

Были ли отделены фиксированные и текущие оформленные тенденции? Были ли недостаточные подгруппы опубликованы как окончательные результаты? Правильный ли статус качества выборки? Были ли дополняющие панели заменены основной панелью? Есть ли четко определенный ответственный за регистрацию выборки?

57. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Почему бы нам не давать каждой стране людей напрямую в соответствии с пропорцией населения?»

Это сильное допущение для стран с большим и средним населением. В очень маленьких странах исходная квота может быть меньше одного человека. В этом случае:

- пул маленьких стран,
- вероятностный отбор стран,
- отдельная панель наблюдателей

может быть более точной.

Возражение 2— «Разве справедливее не было бы дать каждой стране хотя бы одного человека?»

Это может быть полезно с точки зрения географического охвата. Но это несоразмерно с точки зрения оценки численности населения. Правильное решение — отдельно обозначить этого одного человека как наблюдателя от страны.

Возражение 3— «Разве то, что страна не получает ни одного человека на основной панели, полностью её исключает?»

Это может быть проблемой, если страна получает нулевую квоту как отдельный детерминированный слой. Положительная вероятность отбора может быть предоставлена в пуле малых стран. Кроме того, Панель наблюдателей может поддерживать наблюдательную видимость страны.

Возражение 4— «Стирает ли объединение малых стран культурные различия?»

Это может быть удалено, если используется для генерации оценки страны. Для основной глобальной оценки пул управляет вероятностью выбора. Идентификатор страны сохраняется в записи наблюдения. Можно проводить отдельную выборку для Наблюдателей и конкретной страны.

Возражение 5— «Почему мы не включаем пользователей Наблюдателя в 1,000?»

Их можно включить. Но тогда выборка людей 1,000 больше не будет пропорциональна чистой популяции. Потребуется взвешивание, и эффективный размер выборки основного показателя может уменьшиться. В исходной архитектуре наблюдения 1,000 распределяются по Панели населения; наблюдения Страны-наблюдателя добавляются к этому.

Возражение 6— «Разве общая стоимость полевых работ не станет очень высокой?»

Может быть. Стандарт не должен создавать искусственную точность, скрывая стоимость. При меньшем бюджете:

- меньше ИИ-продуктов,
- меньше языков,
- меньше пилотного проекта,
- большая неопределённость

может использоваться. Правильный статус должен быть четко указан.

Возражение 7— «Почему мы собираем 1,093 или больше наблюдений, когда существует имя GEO-1000?»

1,000 является допустимой целью наблюдения для основной панели населения. Панели справедливости и охвата могут быть добавлены к основной цели. Это не нарушает название стандарта. Это делает видимым фактический общий объем поля измерения.

Возражение 8— «Почему чрезмерная выборка одного языка искажает глобальный результат?»

Чрезмерно выбранный язык выглядит большим, потому что он представлен в необработанных ответах. При возврате к весу населения глобальный результат не искажается. Отдельный результат по языку становится более точным.

Возражение 9— «Веса очень сложные; разве необработанное среднее не проще понять?»

Необработанное среднее может быть значимым, если оно пропорционально выборке населения. Чрезмерная выборка, ненадежные ответы или различия в охвате могут привести к тому, что необработанное среднее отвечает не на тот вопрос. Сложность нельзя скрыть, если она возникает из дизайна выборки.

Возражение 10— «Если существует 1,000 действительных наблюдений, почему эффективная выборка должна быть меньше?»

Потому что некоторые наблюдения могут иметь очень высокий вес, в то время как другие могут иметь очень низкий вес. Записи от одной и той же панели или пользователя могут быть похожи друг на друга. Сырой показатель — это количество физических записей. Эффективная выборка приближается к объему статистической информации.

Возражение 11— «Разве не будет справедливее, если одни и те же люди протестируют все ИИ-продукты?»

Снижает разницу между пользователями. Однако:

- это может сузить доступность продукта,
- это может увеличить нагрузку на задачи,

это может создать перекрестное влияние между ответами. Это можно контролировать с помощью сопоставленного или сбалансированного дизайна неполных блоков.

Возражение 12— «Если слой не достигает цели, может ли глобальный рейтинг никогда не публиковаться?»

Не в каждом случае. Недостаток оценивается с точки зрения:

- население вес,
- риск,
- причина недостаточности,
- остающееся покрытие

учтено. Балл:

- с узким покрытием,
- с большой неопределенностью,
- С предупреждением НЕДОСТАТОЧНОЕ ПОКРЫТИЕ

его можно публиковать. Недостаток нельзя скрыть.

Апелляция 13— «Если данные по населению изменятся, нужно ли обновлять все квоты в каждой волне?»

Обновление может быть необходимо для текущего мгновенного балла. Фиксированные начальные веса также могут поддерживаться для сравнения тенденций. Оба результата должны публиковаться отдельно.

Апелляция 14— «Если мы не можем создать вероятностную мировую панель, можно ли не применять NOMOS?»

Это применимо. Однако статус качества образца должен быть правильно зафиксирован. Панель добровольцев, калиброванная относительно населения, может быть ценным инструментом. Ее нельзя представлять как полностью вероятностную выборку.

ОБЩЕЕ ПОЛОЖЕНИЕ ГЛАВЫ 60

Мощь GEO-1000 не возникает только из числа 1,000. Из 1,000 человек:

- из какой они были популяции,
- с какой вероятностью они были выбраны,
- какая страна и язык они представляли,
- сколько раз один и тот же человек учитывался,
- какие наблюдения считались действительными,
- как обрабатывались маленькие страны,
- как взвешивались группы с повышенной выборкой,
- сколько человек не завершили задание,
- кто был заменен на основе результатов

Авторитет образца возникает из этих проектных решений. Дефектный образец может давать ответы 100,000 и при этом говорить о неправильной популяции. Надежный образец может содержать меньше ответов, но поддерживать более достоверный вывод в пределах ясно обозначенных ограничений. Население большой страны должно иметь значительный вес в глобальном человеческом опыте, но пользователей нельзя выдумывать там, где продукт недоступен. Глобальный вес маленькой страны должен оставаться небольшим, но страна не обязана быть невидимой. Язык может составлять небольшую долю мирового населения и при этом иметь существенную измеримую ошибку. Один человек может использовать три языка и пять продуктов; это не создает пятнадцать человек. Ответ ИИ может быть неверным, не делая участника недействительным. Неверный ответ — это то, что аудит пытается измерить.

Замена его на нового пользователя не очищает измерение. Это стирает реальность. Взвешивание чудесным образом не превращает неправильную панель в мировое население. Однако правильно разработанный оверсемплинг может сделать подгруппы видимыми без искажения глобальной оценки. Поэтому шестой закон измерения NOMOS таков:

> *Выборка должна создаваться не для того, чтобы получить результат, а для того, чтобы наблюдать результат честно.*

Его седьмой закон таков:

> *Видимость маленькой страны не должна увеличивать вес её населения.*

Его восьмой закон таков:

> *Чрезмерное наблюдение не означает чрезмерный вес населения.*

Девятый закон звучит следующим образом:

> *Неправильный ответ ІІІ1 не является дефектом данных; если он наблюдается по действующему протоколу, это вопрос надзора.*

Десятый закон таков:

Уверенность в тысяче наблюдений не возникает из числа тысяча; она возникает из воспроизводимости отбора и взвешивания.

Порядок главы 6 NOMOS

Не просто говори мне: «Мы опросили 1,000 человек.»

Покажите, из какого миллиарда людей были выбраны эти 1,000 и с какой вероятностью.

Не игнорируйте страну только потому, что она составляет половину человека. / Но также не утверждайте, что измерили всю страну, только потому что опробовали одного человека.

Выбирайте маленькие страны вероятностно в соответствующем пуле. / Если требуется видимость, создайте отдельную панель наблюдателя.

Вы можете делать оверсэмплинг языков с низким ресурсом. / Но не делайте их больше в общем балле, чем их население.

Не размножайте человека так же, как аккаунты, языки, сессии и продукты III.

Используйте резервный вариант, когда участник не отвечает. / Не меняйте своего пользователя, когда III дает неправильный ответ.

Определите правильность, не видя правильности вашего ответа.

Если квота отсутствует, объясните это. / Не привлекайте человека из другой страны, чтобы закрыть пробел.

Не изменяйте долю, вес, пула стран или целевую выборку после того, как увидели результат.

Показывайте сырое число. / Показывайте эффективную выборку. / Показывайте, кто получил большой вес, а кто — малый.

Сначала заблокируйте население. / Затем настройте слои. / Затем рассчитайте квоты. / Затем выберите резервные варианты. / Затем запишите веса. / И только после этого отправляйте первый запрос.

Заключительное предложение главы

GEO-1000— это не метод компиляции тысячи ответов; это протокол, позволяющий сделать измеримым представление каждого человека в целевой популяции с учетом вероятности отбора и корректного взвешивания.

Нормативное ядро

Для каждого продукта ИИ и основной волны измерений Панели населения GEO-1000 СЛЕДУЕТ стремиться к охвату как минимум 1,000 действительных наблюдений продукта, если явно не обоснована другая цель. Распределение выборки ДОЛЖНО быть спроектировано, версионировано и зафиксировано до того, как будут получены ответы ИИ. Основное распределение ДОЛЖНО основываться на целевой популяции, соответствующей продукту, а не на общей численности жителей, количестве приглашений, учетных записях, сессиях или доступности по удобству. Дробные распределения ДОЛЖНЫ быть преобразованы в целочисленные квоты с использованием заранее объявленного, воспроизводимого метода, слепого к результатам. Популяциям с положительной массой НЕЛЬЗЯ детерминированно присваивать нулевую вероятность отбора, когда аудит утверждает выводы для всей целевой популяции. Малые популяции МОГУТ быть представлены через обоснованные объединённые страты и выборку с вероятностью, пропорциональной размеру. Наблюдения Country Observer, языковой справедливостью и доступностью МОГУТ дополнять основную панель, но ДОЛЖНЫ оставаться отдельно маркированными и соответственно перенормированными. Избыточно представленным группам НЕЛЬЗЯ назначать их исходную долю выборки в качестве веса популяции. Замена участников ДОЛЖНА основываться только на заранее объявленных условиях набора или условиях допустимости протокола. Некорректные, отрицательные, отказавшиеся или критические результаты ИИ НЕЛЬЗЯ использовать для инициирования замены. Лица, учетные записи, сеансы, языки и наблюдения за продуктами ИИ ДОЛЖНЫ оставаться отдельными единицами. Дизайн-веса, корректировки на неответы, калибровка, усечение, кластеризация, парные наблюдения и эффективный размер выборки ДОЛЖНЫ быть задокументированы. Любое изменение квот, страт, пулов, весов, исключений или целей выборки после получения результатов ДОЛЖНО создавать новую версию анализа и сохранять исходную запись.

ГЛАВА 07

Панели странового наблюдения и языковой справедливости

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K04; K06; K07; K20

Граница главы

Раздел 5 установил следующее положение:

> *Для каждого AI продукта необходимо отдельно определять популяцию пользователей, подходящих для продукта.*

Раздел 6 преобразовал эту популяцию в выборку:

> *Основная панель популяции должна предоставлять заранее определенные вероятности выбора и веса для людей в целевой популяции до того, как будут видны результаты.*

Теперь нам нужно решить две разные задачи справедливости. Первая проблема:

> *Взвешивание по популяции может сделать очень маленькие страны невидимыми в основной выборке.*

Вторая проблема:

> *Среднее значение по мировой популяции может скрывать серьезные искажения представления в языках с низкими ресурсами или низкой численностью населения.*

Чистая выборка по популяции сильна для измерения среднего человеческого опыта в мировой популяции. Однако та же выборка:

- не может наблюдать каждую страну индивидуально,
- не может предоставить достаточное количество выборок для каждого языка
- не удается создать оценку страны для небольших стран,

может не выявлять редкие, но существенные ошибки на языках с ограниченными ресурсами. Поэтому GEO-1000 не может состоять только из одной панели. Вместе с основной панелью населения необходимы две дополнительные структуры:

Панель наблюдателей страны

Стандартная метка панели:

Панель наблюдателей страны

и:

панель языковой справедливости

Стандартная метка панели:

панель языковой справедливости

Панель наблюдателей за страной отвечает на следующий вопрос:

> *В странах, которые недостаточно представлены в основной панели населения, есть ли хотя бы реальные наблюдения пользователей, проверка доступа или ранние предупреждающие сигналы?*

С другой стороны, Панель языковой справедливости отвечает на следующий вопрос:

> *Даже если глобальный средний показатель кажется высоким, представлены ли пользователи на определенных языках неправильно, не полностью или рискованно?*

Эти две панели не заменяют основную Панель Населения. Они дополняют её. Предоставление наблюдателю страны возможности наблюдать за страной не измеряет результат этой страны. Чрезмерная выборка языка не делает этот язык более распространённым в мировом населении. Задачи этого раздела:

- видимость страны по сравнению с весом населения,
- разделение глобального среднего с учётом языковой справедливости,
- отделение оценки численности населения от раннего предупреждения,
- разделение статистической достаточности от объёма наблюдений

друг от друга. Этот раздел ещё не:

- определил, как промпты должны переводиться на все языки,
- провёл обратный перевод и тест на семантическое соответствие,
- установил детальный механизм отбора участников,
- описал, как атомарные утверждения будут оцениваться арбитрами,
- установил окончательные пороги соответствия

точно. Эти вопросы будут дополнительно регулируются в следующих разделах. Основной вопрос Раздела 7:

> *Как мы измеряем, создавая глобальный рейтинг, не искусственно завышая показатели маленьких стран и языков, но при этом не оставляя их полностью незаметными?*

NOMOS Вызов

Вы создаете панель населения 1,000 для продукта на базе ИИ. В результате выборки пропорциональной населению:

- сотни пользователей из нескольких крупных стран,
- десятки пользователей из некоторых средних стран,
- нет пользователей из большинства маленьких стран

прибыло. Глобальный рейтинг:

94

Это было рассчитано таким образом. Затем вы сказали: «Этот продукт ИИ представляет бренд по всему миру с точностью 94 процентов». Однако в 70 маленьких странах нет ни одного реального наблюдения пользователей. Эти страны могут составлять небольшую часть мирового населения. Тем не менее, остаются следующие вопросы без ответа: действительно ли продукт доступен? Работает ли намерение на правильном языке? Система путает одну страну с другой? Местная компания путается с глобальным брендом? Неверны ли цена, сервис и юридический контекст? Не удается ли системе отвечать? Есть ли редкие критические ошибки?

Глобальный показатель населения не отвечает на эти вопросы. Теперь представьте, что вы добавили одного человека в каждую страну. Вы получили 70 дополнительных наблюдений.

Вы добавили каждую из этих стран к основному баллу с весом 1/1,070, делая маленькие страны многократно больше их истинной доли населения. Географическая видимость увеличилась, но оценка населения была искажена. Затем один пользователь в одной стране получил неправильный ответ, и вы заявили: «GEO-балл этой страны равен нулю». Ответ одного человека не является баллом страны. Это единичное событие, наблюдаемое в определённых условиях в этой стране. Теперь рассмотрим язык. Результат для английского — 98 процентов, для немецкого — 96 процентов, для турецкого — 91 процентов. Язык с низким ресурсом имеет всего восемь пользователей и сырое значение 50 процентов. Вы говорите: «Восьми человек недостаточно; уберите язык из отчета». Глобальный рейтинг растет, но опыт пользователей на этом языке исчезает.

В другом подходе вы предоставляете дополнительные 200 пользователей для языка с низкими ресурсами. Среднее значение выборки снижается. Затем вы говорите: "Глобальный показатель снизился до 84." Однако фактическая доля населения, говорящая на этом языке, может составлять 2 процентов. Чрезмерная выборка увеличивает статистическую мощность. Она не увеличивает вес в популяции. Первое правило этого раздела следующее:

> *Наблюдаться — не значит иметь вес.*

Второе положение гласит:

> *Отдельное наблюдение по стране — это не оценка страны.*

Её третье положение гласит:

> *Справедливость по языкам не означает присваивание равного веса всем языкам в глобальном показателе.*

Четвертая статья гласит:

> *Высокое глобальное среднее не устраняет серьезного ухудшения в конкретном языке.*

Пятая его статья гласит:

> *Дополнительные наблюдения, сделанные для увеличения охвата, нельзя добавлять к основному показателю без перерасчета веса в соответствии с истинной популяцией.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела заключается в создании дополнительных панелей, которые повышают видимость стран и языков без нарушения представления статистической популяции основной Панели Населения. В разделе нормализуются следующие различия:

- Географическая видимость с учетом веса популяции
- Наблюдение за страной с оценкой страны
- Обнаружение одиночного события с оценкой распространенности
- Основная Панель Населения с Панелью Наблюдателей Стран
- Панель Наблюдателей Стран с валидационным исследованием, специфичным для страны
- Покрытие стран с покрытием популяции
- Покрытие числа языков с учетом численности носителей
- Справедливость по языку с равным глобальным взвешиванием языков
- Глобальная доля языка с чрезмерной выборкой

- Родной язык и язык запроса
- Язык и локаль
- Язык и система письма
- Поддерживаемый язык и язык, который может самостоятельно генерировать ответы
- Дефект перевода промпта и языковой дефект ИИ-системы
- Недостаток ресурсов и дефект представления ИИ
- Недостаточность арбитров и сбой продукта
- Законное местное различие и несправедливое языковое неравенство
- Глобальный средний показатель и база подгруппы
- Охват доступа по странам и точность представления стран
- Сигнал наблюдателя и окончательное решение о соответствии
- Одноволновой цикл наблюдения с многоволновым географическим покрытием

В конце этого раздела каждый аудит GEO-1000 должен быть способен дать четкие ответы на следующие вопросы:

> *В каких странах есть хотя бы одно фактическое наблюдение?*
>
> *В каких странах есть только наблюдение по раннему предупреждению, а в каких — статистические прогнозы?*
>
> *Какие языки были измерены, какие достигли достаточного объема выборки, а какие не были протестированы?*
>
> *Если языки с низкими ресурсами были избыточно выбраны для выборки, как это было учтено в глобальной оценке?*
>
> *Является ли источник плохого результата на языке действительно сгенерированным ИИ, или это связано с промптом, ресурсами, локализацией или проблемами с экспертизой?*

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

> *GEO-1000 должен управлять основной оценкой численности с помощью Панели населения; наблюдением за небольшими странами с помощью Панели наблюдателей стран; идентификацией языкового неравенства и языков с низкими ресурсами отдельно с помощью Панели языковой справедливости.*

Эти три структуры:

- отличаются по целям,
- отличаются по целевым выборкам,
- отличаются по весам,
- отличаются по публичным утверждениям

имеют. Наблюдение одной панели не может быть преобразовано в значение другой панели без объяснения. Структура по умолчанию:

Панель	Основная цель	Результаты, которые он может дать	Результаты, которые он не может дать самостоятельно
Панель населения	Глобальная оценка с учетом численности населения	Глобальный и слоистый показатель населения	Надежный показатель для каждой небольшой страны
Панель наблюдателей страны	Географическое наблюдение и раннее предупреждение	Покрытие наблюдения отдельных стран, отдельное событие, сигнал доступности	Соотношение населения стран

Панель	Основная цель	Результаты, которые он может дать	Результаты, которые он не может дать самостоятельно
панель языковой справедливости	Диагностика и измерение неравенства на основе языка	Языковые показатели, языковой минимум, разрыв справедливости	Автоматическое равномерное взвешивание языков в мировом населении

3. ЧТО ТАКОЕ ПАНЕЛЬ СТРАНОВЫХ НАБЛЮДАТЕЛЕЙ?

Панель страновых наблюдателей в странах, которые недопредставлены или недостаточно представлены в основной панели населения:

- доступ к продукту,
- Применимость промпта
- базовый анализ сущностей,
- поведение языка и локали,
- ранний сигнал критической ошибки

является дополнительной аудиторской панелью, созданной для наблюдения. Основная задача Панели Наблюдателей за Странами не заключается в выставлении оценки.

> *Она заключается в том, чтобы впервые сделать невидимую страну видимой.*

Наблюдение Country Observer может показать: «В указанной стране, для указанного ИИ-продукта пользовательский интерфейс, языка и временного условия, этот результат наблюдался хотя бы один раз». Оно не может показать: «X% пользователей в этой стране видят тот же результат».

4. Единица Панели Наблюдателей за Странами

Базовой единицей наблюдения Панели Наблюдателей за Странами является не только страна. Предпочтительная единица:

```
S_{c,a,l,s,w}
```

где:

- c: страна или юрисдикция
- a: Продукт ИИ и пользовательский интерфейс
- l: язык, система письма и локаль
- s: условие сессии или плана
- w: волна измерений

Это различие необходимо. В одной и той же стране:

- более одного языка,
- разные планы,
- веб- и мобильные интерфейсы,
- разные юридические или технические условия доступа

могут существовать. Добавление пользователя в страну не означает мониторинг всех языковых и пользовательских интерфейсов страны.

5. ОБЯЗАННОСТИ ПАНЕЛИ НАБЛЮДАТЕЛЯ СТРАНЫ

5.1. Проверка доступа

Хотя продукт, согласно официальной регистрации, доступен в стране, фактический пользователь:

- может не иметь возможности войти в систему,
- может не иметь возможности достичь указанной интерфейсы,

Может столкнуться с препятствием в плане или проверке. Наблюдатель страны тестирует официальную запись доступа на полевом уровне.

5.2. Базовая проверка личности

Сущность, контролирующая продукт ИИ:

- правильная компания,
- правильный бренд,
- правильные отношения с страной

Может ли он анализировать внутри?

5.3. Обнаружение отклонений локали

В разных локалях одной и той же страны на одном языке:

- цена,
- компания,
- услуги,
- закон,
- совет

Меняется ли это? Если это меняется, может ли эта разница быть финансовой и объяснимой?

5.4. Сигнал редкой тяжёлой ошибки

В ответе одного наблюдателя:

- неправильная лицензия,
- неправильное юридическое лицо,
- опасная информация о здоровье или безопасности,
- тяжелое слияние идентичности

Это можно наблюдать. Одинарное событие не указывает на распространенность в стране. Однако оно может инициировать проверку.

5.5. Новая гипотеза исследования

Наблюдательная панель может задать вопрос: «Является ли это событие единичным, или оно повторяющийся шаблон среди пользователей в стране?» Этот вопрос проверяется с помощью целевого дополнительного образца.

5.6. Запись о географическом охвате

Она показывает, в скольких подходящих странах продукт имеет реальные наблюдения пользователей. Это не количество отдельных оценок стран. Это наблюдательное покрытие.

6. ЧТО НАБЛЮДАТЕЛЬНАЯ ПАНЕЛЬ СТРАНЫ НЕ МОЖЕТ СДЕЛАТЬ

Только наблюдательная панель страны:

- надежный счет страны,
- уровень населения страны,
- рейтинг страны,
- решение о соответствии страны,
- превосходство модели на основе страны

не может быть получено. Если в стране есть только один наблюдатель и ответ верный: «Оценка страны 100» нельзя утверждать. Если неверный: «Оценка страны равна нулю» нельзя утверждать. Верный статус:

OBSERVED_PASS

Наблюдаемая проблема

OBSERVED_CRITICAL_CANDIDATE

СБОЙ ДОСТУПА

НЕДОСТАТОЧНО ДЛЯ ОЦЕНКИ ПО СТРАНЕ

являются наблюдаемыми записями.

7. РЕЕСТР ПРАВОМОЧНЫХ СТРАН

Для каждого ИИ-продукта должен быть создан реестр право¬мочных стран. Каждая страна может иметь один из следующих статусов:

CE-0— НЕИЗВЕСТНО

Статус доступа к продукту и аудиту не удалось определить.

CE-1— ВНЕ Естественный охват

Продукт остается вне официальной или легитимной зоны использования. Он не учитывается в оценке точности представления страны. Показывается как пробел в области доступа.

CE-2— ОГРАНИЧЕННЫЙ ИЛИ СПЕЦИАЛЬНЫЙ ДОСТУП

Только:

- учебное заведение,
- приглашениями,
- план,
- устройство,
- тип пользователя

доступ доступен. Может потребоваться отдельная панель.

CE-3— ПОДХОДИТ ДЛЯ НАБЛЮДЕНИЯ Country Observer

Продукт подходит для наблюдения Country Observer с точки зрения соответствующего пользовательский интерфейс и как минимум одного языка аудита.

CE-4— ПОДХОДИТ ДЛЯ ОЦЕНКИ НАСЕЛЕНИЯ

Имеется доля населения, подходящая для продукта, в основной панели населения страны.

CE-5— ПОДХОДИТ ДЛЯ ОЦЕНКИ СТРАНЫ

Можно получить оценку страны при достаточной выборке, эффективной выборке, охвате языков и условиях качества. Страна может быть CE-3, но не CE-5. Это означает, что страна наблюдается, но надежную оценку страны получить не удалось.

8. ЛЕСТНИЦА МЕЖДУ НАБЛЮДЕНИЕМ Country Observer И ОЦЕНКОЙ СТРАНЫ

CS-0— НЕ НАБЛЮДАЕТСЯ

Для соответствующей страны нет действительного наблюдения наблюдателя страны.

CS-1— ОДИНОЧНОЕ наблюдение наблюдателя страны

Есть одно действительное наблюдение «страна–продукт–язык». Генерирует только запись события.

CS-2— ПОВТОРНОЕ наблюдение наблюдателя страны

Есть несколько наблюдений из разных волн или пользователей. Может вызвать подозрение в шаблоне. Недостаточно для расчета показателя страны.

CS-3— ЭКСПЛОРАТОРНАЯ ВЫБОРКА СТРАНЫ

Достигнута предельная группа эксплороторной выборки в разделе 6. Можно дать эксплороторный результат с большой неопределенностью.

CS-4— ПРОВИЗОРНАЯ ОЦЕНКА СТРАНЫ

Можно получить провизорную оценку страны при достаточном и сбалансированном отборе проб.

CS-5— КАНДИДАТ НА ОЦЕНКУ СТРАНЫ

Достаточное количество активных образцов, охват по языкам/локалям, повторные волны и контроль качества обеспечены. Он может быть кандидатом на оценку по стране. Точные пороговые значения следует откалибровать с помощью пилотного исследования и внешнего обзора.

9. ЦИКЛ НАБЛЮДАТЕЛЯ

Наблюдение каждой волны во всех подходящих странах может быть невозможным с точки зрения затрат и операций. Поэтому может использоваться Цикл Наблюдателя. Волны измерений:

```
w=1,2,…,K
```

и подходящие страны: пусть c∈CS. Назначение наблюдателя для страны c в волне w:

```
z_{c,w}= 1(если страна будет наблюдаться в этой волне); в противном случае z_{c,w}= 0
```

может быть представлено следующим образом. Условие базового цикла:

```
\sum_{w=1}^K z_{c,w}\geq 1
```

должно соблюдаться. То есть каждая подходящая страна должна быть наблюдаема хотя бы один раз в пределах определённого цикла наблюдения. Это положение:

- не означает, что каждая страна будет наблюдаться в каждой волне,
- или что одно наблюдение будет составлять оценку страны

Ни одно из условий не следует из требования цикла.

10. КЛАССЫ ЧАСТОТ НАБЛЮДЕНИЯ

Страны могут быть разделены на заранее определённые классы частот.

SF-0— ВНЕ ДОСТУПА

Нет доступа к продукту. Наблюдение представителя пользователя не проводится. Статус доступа контролируется.

SF-1— ОДИН РАЗ ЗА ЦИКЛ НАБЛЮДЕНИЯ

Страны с низким риском, небольшие страны или страны, которые уже имеют достаточное косвенное покрытие в Панели населения.

SF-2— ПЕРИОДИЧЕСКОЕ НАБЛЮДЕНИЕ

Наблюдается снова через равные промежутки времени.

SF-3— КАЖДАЯ ОСНОВНАЯ ВОЛНА

Может наблюдаться в каждой основной волне по следующим причинам:

- высокая численность пользователей,
- материальный целевой рынок,
- предыдущая серьезная ошибка,
- быстрый доступ или изменения в регулировании,

зона обслуживания с высоким риском.

SF-4— МОНИТОРИНГ ИНЦИДЕНТОВ

Усиленное наблюдение из-за подтвержденного критического события или непрерывного нарушения представления. Класс частоты не может быть изменен ретроспективно после рассмотрения результатов. Новое событие может создать новую версию для следующей волны.

11. ПРИОРИТЕТ НАБЛЮДАТЕЛЯ

Частота активности наблюдателя страны не должна определяться исключительно на основе численности населения. По мере необходимости могут учитываться следующие области:

- Потенциальная для продукта аудитория

- Фактическое покрытие услуги или рынка объекта
- Предыдущие записи о неправильном представлении
- Влияние на пользователей с высоким риском
- Изменения в доступе к продукту
- Уникальность языка и локали
- Использование языков с ограниченными ресурсами
- Юридические или операционные различия
- Время, прошедшее с момента последнего наблюдения

Эти поля должны использоваться в соответствии с заранее определенным правилом управления. Подход «Результат этой страны плох, давайте измерять меньше» запрещен.

12. Метрики покрытия наблюдателем страны

Покрытие страны не является одним числом. Должны использоваться как минимум три отдельных метрики.

12.1. Покрытие наблюдения страны

Доля стран, имеющих как минимум одного действительного наблюдателя или наблюдение за населением.

```
COC_count= |{c ∈ C_S: n_c^obs > 0}| / |C_S|
```

Эта метрика показывает покрытие с точки зрения числа стран.

12.2. Покрытие наблюдения страны с учетом населения

Доля подходящей популяции в странах с хотя бы одним действительным наблюдением по отношению к общей целевой популяции.

```
COC_pop= [Σ_{c:n_c^obs>0} N_c] / [Σ_{c∈C_S} N_c]
```

Этот показатель показывает, наблюдаются ли большие популяции.

12.3. Покрытие прогнозирования по странам

Это доля стран, достигших достаточной выборки по стране.

```
CEC= |{c: n_{c,eff}≥ m_c}| / |C_S|
```

Здесь mc — это заранее определённый порог адекватности для прогноза по стране. Аудит:

- 100 процент покрытия наблюдений по странам,
- 99 процент покрытия наблюдений по населению,
- может иметь только 10 процент покрытия прогноза по странам

Эти результаты нельзя преобразовать друг в друга.

13. СИГНАЛ ПРЕДУПРЕЖДЕНИЯ НАБЛЮДАТЕЛЯ

Когда в наблюдении наблюдателя обнаруживается материальная проблема, создается сигнал тревоги наблюдателя. Типы сигналов тревоги:

SA-1— АНОМАЛИЯ ДОСТУПА

Официальная запись доступа противоречит фактическому доступу пользователя.

SA-2— АНОМАЛИЯ СУБЪЕКТА

Субъект объединяется с другой компанией, продуктом или операцией в стране.

SA-3— АНОМАЛИЯ ЯЗЫКА ИЛИ ЛОКАЛИЗАЦИИ

Происходит неверный язык, неправильный контекст страны или очевидное отклонение локали.

SA-4— АНОМАЛИЯ ВАЖНОГО ФАКТА

Материальная цена, услуга, мощность, партнерство или время указаны неверно.

SA-5— КАНДИДАТ НА КРИТИЧЕСКИЙ ИНЦИДЕНТ

Существуют проблемы со здоровьем, законом, финансами, безопасностью, лицензированием или серьезная ошибка идентичности.

SA-6— АНОМАЛИЯ ССЫЛКИ ИЛИ промпта

Источник проблемы не является ответом ИИ:

- запрос,
- перевод,
- официальный источник,
- разбирательство

может быть виноват. Тревога не является автоматической ошибкой страны. Это начало проверки и эскалации.

14. ПРОТОКОЛ ЭСКАЛАЦИИ НАБЛЮДАТЕЛЯ

Шаг 1— Проверка достоверности

Соответствует ли наблюдение действительно протоколу? Верная страна Верный продукт Верная промпт Верный язык Первоначальный подходящий результат Все доказательства проверены.

Шаг 2— Перекодирование инцидента

Проблема:

- Продукт ИИ,
- реестровая запись,
- запрос,
- участник,
- доступ

К какому источнику это может принадлежать?

Шаг 3— Независимое повторение

Прежде чем выбрать результат, выполняется проверка с заранее заданными новыми пользователями или сессиями.

Шаг 4— Целевое выборочное обследование

Если событие продолжается, может применяться выборка по странам. Размер выборки:

- тяжесть события,
- ожидаемая распространенность,
- население страны,
- языковое разнообразие

заранее определяются.

Шаг 5— Разделение решений

Следующие результаты сохраняются отдельно:

- Одно подтвержденное событие
- Повторяющийся шаблон
- Уровень по стране
- Критический эффект применимости

Одно критическое событие может влиять на проверку соответствия даже без оценки распространенности.

15. ЧТО ТАКОЕ ЯЗЫКОВАЯ СПРАВЕДЛИВОСТЬ?

Языковая справедливость:

> *Речь идет не о том, чтобы все языки получали точно одинаковый ответ слово в слово.*

Это означает:

> *В семантически эквивалентных и материально сопоставимых пользовательских вопросах, кроме реальных локальных различий, пользователи не должны систематически получать представления, которые более ошибочны, более неполные, меньше обоснованы или более рискованны из-за их языка.*

Языковая справедливость не стремится к этому равенству:

```
R_Turkish= R_English= R_Japanese
```

Слово в слово. Она ищет это:

> *Сопоставимое представление с точки зрения идентичности, материальной реальности, границ, времени, статуса источника и пригодности для пользователя.*

16. ЯЗЫКОВАЯ РАЗЛИЧНОСТЬ НЕ ВСЕГДА ЯВЛЯЕТСЯ НЕДОБРОСОВЕСТНОСТЬЮ

Языковые результаты могут законно различаться по следующим причинам: Объем местных услуг различен. Цены и условия контрактов варьируются в зависимости от страны. Юридические знания зависят от юрисдикции. Намерения пользователя различаются. промпты не являются семантически идентичными. Официальные источники содержат разный контент на разных языках. Продукт AI официально не поддерживает соответствующий язык. Каждое из этих различий следует фиксировать. Объяснимые местные различия не следует отмечать как несправедливость. Однако, даже если реальность одинакова, в языке с ограниченными ресурсами:

- идентичность компании повреждена,
- границы сняты,
- добавлена неправильная лицензия,
- ответ постоянно дается на другом языке

это может быть проблема с языковым представлением.

17. ЯЗЫКОВОЕ ПОДРАЗДЕЛЕНИЕ

В оценке языка указание только общего названия языка может быть недостаточным. Предпочтительная языковая единица:

```
L=(λ,σ,ℓ,d)
```

и может рассматриваться как. Здесь:

- λ: язык
- σ: письменность
- ℓ: локаль или контекст страны
- d: задача или область

Например, один и тот же язык:

- в другой письменной системе,
- в другой стране,
- в юридической или медицинской области

может потребовать отдельной измерительной ячейки. Не все пользователи языка представляют собой единую однородную группу.

18. СТАТУСЫ ПОДДЕРЖКИ ЯЗЫКА

Каждый ИИ-продукт может использовать один из следующих статусов для языка и пользовательский интерфейс:

LS-0— НЕИЗВЕСТНО

Статус поддержки не может быть определён.

LS-1— НЕ ПОДДЕРЖИВАЕТСЯ

Продукт официально не поддерживает рассматриваемый язык. Он не оценивается в рамках оценки основного поддерживаемого языка. Может проводиться экспериментальное тестирование.

LS-2— НАБЛЮДАЕТСЯ, НО НЕ ПОДДЕРЖИВАЕТСЯ

Продукт может генерировать ответы. Однако официальной поддержки или гарантии качества нет. Сообщается как отдельный экспериментальный результат.

LS-3— ОГРАНИЧЕННАЯ ПОДДЕРЖКА

Поддержка существует для конкретных задач, интерфейсов или функций.

LS-4— ОФИЦИАЛЬНО ПОДДЕРЖИВАЕТСЯ

Язык продукта в целом заявляется как поддерживаемый.

LS-5— ДОПУСКАЕТСЯ К АУДИТУ

Язык:

- официальная поддержка,
- эквивалентность промптов,
- местный арбитр,
- ссылочная компетентность,
- допустимое захватывание

полностью соответствует требованиям аудита. Само наличие официальной поддержки не является LS-5.

19. ЯЗЫК С НИЗКИМИ РЕСУРСАМИ

Термин «язык с низкими ресурсами» не описывает ценность языка или качество его носителей. В отдельных технологических и аудиторских контекстах он указывает на один или несколько из следующих недостатков:

- Ограниченное цифровое содержание
- Ограниченный локализованный официальный источник
- Ограниченный набор данных для оценки
- Ограниченный доступ к оценщикам
- Низкая поддержка продукта
- Ограниченные терминологические ресурсы
- Низкое качество машинного перевода
- Ограниченный доступ к пользовательской панели

Язык может быть с низкими ресурсами в одной области и сильным в другой. Например, хотя он имеет сильные ресурсы в повседневном общении:

- закон,
- медицина
- конкретная техническая область

может иметь ограниченные ресурсы. Следовательно, уровень языковых ресурсов:

- продукта,
- задача,
- время,
- область

должны быть определены вместе.

20. ЗАДАЧИ ПАНЕЛИ ПО ЯЗЫКОВОМУ СПРАВЕДЛИВОСТИ

20.1. Увеличение выборки для языков с низким уровнем ресурсов

Предоставляет дополнительных пользователей для языков, которые не получают достаточное количество наблюдений в основной панели населения.

20.2. Измерение распределения материальных ошибок по языкам

Не только общая точность:

- идентичность,
- объем охвата,
- время,
- источник,
- рекомендация,
- Критическая ошибка

исследует результаты отдельно.

20.3. Проверка языка ответа

Отвечает ли ИИ в соответствии с языком запроса? Если он переключается на другой язык, какова причина?

20.4. Исследование различий в языке и локали

Один и тот же язык несет разную локальную реальность в разных странах? Способен ли ИИ правильно учитывать это различие?

20.5. Исследование паритета источника и официального представления

Несут ли официальные записи существования одинаковую материальную реальность на всех языках? Низкий результат ИИ может возникнуть из-за языкового разрыва в самом источнике.

20.6. Тестирование эквивалентности промптов

Действительно ли переводы передают один и тот же пользовательский замысел?

20.7. Тестирование равенства арбитров

Могут ли арбитры на разных языках применять один и тот же нормативный стандарт?

21. ЧЕТЫРЕ ВОРОТА ЯЗЫКОВОЙ СПРАВЕДЛИВОСТИ

Языковое достижение должно оцениваться через четыре ворота, прежде чем приписывать его языковой производительности ИИ-продукта.

Ворота 1— Эквивалентность замысла

Являются ли промпты семантически сопоставимыми? Если это не удается:

ОШИБКА_ЭКВИВАЛЕНТНОСТИ_ПОДСКАЗОК

статус указан.

Ворота 2— Эталонное соответствие

Достаточны ли официальные и проверенные записи реальности для соответствующего языка? Если нет:

ОТСУТСТВИЕ_ЯЗЫКОВОЙ_ЭТАЛОННОЙ_БАЗЫ

статус указан.

Ворота 3— Соответствие измерений

Сравнимы ли условия захвата, сессии, интерфейса и пользователя? Если это не удается:

ОШИБКА_СООТВЕТСТВИЯ_ИЗМЕРЕНИЙ

статус указан.

Ворота 4— Соответствие оценивания

Есть ли оценщики на родном языке или с достаточной локальной экспертизой и одинаковыми правилами оценки? Если это не удается:

ОТСУТСТВИЕ_ЯЗЫКОВОЙ_ЭКСПЕРТИЗЫ_ОЦЕНЩИКОВ

назначается. Если только четыре условия не выполнены надлежащим образом: «Продукт на базе ИИ хуже на этом языке» не может быть окончательно установлено. Языковой разрыв по-прежнему является находкой экосистемы GEO. Однако его источник должен быть правильно классифицирован.

22. УСТАНОВЛЕНИЕ ЯЗЫКОВОЙ ВСЕЛЕННОЙ

Фраза «все языки» не может использоваться без ограничений. Каждый аудит должен создать Языковой Реестр в рамках Объема. Языки могут быть выбраны из следующих групп:

22.1. Глобальный набор основных языков

Родные языки, используемые большим числом людей и подходящие для продукта.

22.2. Набор языков рынка сущности

Языки, на которые нацелена проверяемая сущность:

- услуги,
- с точки зрения содержания,
- клиент,
- контракт

и поддержки.

22.3. Набор языков поддержки продукта

Языки, официально поддерживаемые продуктом на базе ИИ.

22.4. Набор языков с ограниченными ресурсами для наблюдения

Языки с ограниченными ресурсами, которые слабо представлены в основной панели населения или несут риск с точки зрения поведения продукта.

22.5. Набор языков высокого риска

Языки, которые оказывают существенное влияние на пользователя в контексте здравоохранения, права, финансов или безопасности.

22.6. Экспериментальный набор языков

Языки, которые продукт официально не поддерживает, но которые тестируются в наблюдательных или исследовательских целях. Эти группы могут пересекаться. Причина включения каждого языка должна быть зафиксирована.

23. РАСПРЕДЕЛЕНИЕ СПРАВЕДЛИВОСТИ ПО ЯЗЫКАМ

Для языка l пусть целевое значение в основной панели населения будет nlP. Дополнительное целевое значение в панели справедливости — nlF. Общее количество собранных наблюдений на языке:

```
n_l^T= n_l^P+ n_l^F
```

получено. Общий вес языка в расчете на мировое население сохраняется как Wl. Нормализованный вес на одно наблюдение:

```
w_{i,l}= W_l/ n_l^T
```

можно рассматривать таким образом. Таким образом, язык подвергается избыточной выборке. Однако он не превышает фактическую долю населения в глобальном балле.

24. ВКЛАД ПАНЕЛИ ПО СПРАВЕДЛИВОСТИ ЯЗЫКА В ГЛОБАЛЬНЫЙ СЧЕТ

Наблюдения за языковой справедливостью могут быть использованы двумя способами.

24.1. Раздельный диагностический анализ

Панель справедливости выдает только оценки по языку и типы ошибок. Основная панель населения не изменяет общий балл. Этот метод проще.

24.2. Объединённый взвешенный анализ

Наблюдения за справедливостью в отношении населения и языка используются вместе. Каждое наблюдение перевзвешивается в соответствии с фактическим весом населения целевого языка. Этот метод может увеличить объем информации. Однако:

- двойной счет
- многоязычный пользователь,
- зависимость панели,
- вариабельность весов

Это следует учитывать. Метод, который будет использоваться, должен быть определен до получения результатов.

25. МЕТРИКИ ПОКРЫТИЯ ЯЗЫКОВ

25.1. Покрытие наблюдения по языкам

Доля языков в области охвата, по которым есть хотя бы одно действительное наблюдение.

```
LOC= |{l: n_l^obs > 0}| / |L_scope|
```

25.2. Покрытие предсказания языков

Доля языков, которые соответствуют требованиям по достаточности выборки и качеству.

```
LEC= |{l: n_{l,eff}≥ m_l}| / |L_scope|
```

25.3. Покрытие языков с учетом численности населения

Доля целевой языковой популяции, которая была адекватно наблюдена. Требуется основной язык аудита или разделение весов пользователей для предотвращения двойного учета многоязычных пользователей.

25.4. Покрытие поддержки продукта

Указывает, на скольких из охваченных языков продукт имеет официальную поддержку. Поддержка языка и точность представления рассматриваются отдельно.

26. Метрики языкового неравенства

Метрики кандидата — Предлагаемые метрики

Точный набор метрик должен быть откалиброван через пилотное и внешнее рассмотрение.

26.1. Результаты по языкам

Для каждой достаточной языковой ячейки рассчитывается θl.

26.2. Разница между наибольшим и наименьшим

```
DG_max= max_l θ_l− min_l θ_l
```

Просто для понимания. Однако чувствительно к маленьким и нестабильным единичным ячейкам.

26.3. Робастная разница процентилей

```
DG_{90−10}= Q_{0.90}(θ_l)− Q_{0.10}(θ_l)
```

Менее чувствительно к экстремальным ячейкам.

26.4. Минимальный уровень языка

Это наименьший результат по языку с достаточной выборкой.

```
LF_min= min_{l∈L_est} θ_l
```

Это значение должно предоставляться вместе с небольшим неопределённым значением для ячейки.

26.5. Надёжный языковой минимум

Это нижний процентиль или среднее значение подгруппы достаточного числа языков. Пример:

```
LF_10= Q_{0.10}(θ_l)
```

26.6. Критический уровень ошибок языка

Для каждого языка:

```
CR_l= допустимый ответ с проверенной критической ошибкой / допустимый языковой ответ
```

Он публикуется отдельно от глобального среднего.

26.7. Доля неизвестного языка

```
U_l= НЕИЗВЕСТНЫЙ или неразрешимый ответ / соответствующее языковое наблюдение
```

Высокая доля неизвестного может быть так же важна, как и низкий показатель.

27. Равный языковой балл и взвешенный по населению балл

Можно получить диагностические результаты, которые оценивают все языки одинаково:

```
G_equal-language= (1/L)Σ_{l=1}^L θ_l
```

Этот результат приближает вопрос: «Средний опыт тестируемого языка». Он не представляет средний человеческий опыт в мировой популяции. Результат с учетом численности населения:

```
G_population-language= Σ_l W_l θ_l
```

Он сохраняет распределение населения. Два показателя различны. Их нельзя использовать взаимозаменяемо. Оба могут быть показаны вместе на карте языковой справедливости.

28. ЯЗЫКОВАЯ СПРАВЕДЛИВОСТЬ НЕ ЯВЛЯЕТСЯ ОБЯЗАТЕЛЬСТВОМ ПО ДОСТИЖЕНИЮ РАВНЫХ РЕЗУЛЬТАТОВ

Иметь одинаковый балл по двум языкам не является автоматическим нарушением. Разница может быть объяснена следующими причинами:

- Фактическая разница местного обслуживания
- Закон, специфичный для страны
- Официальное покрытие продукта поддержкой
- Доступ к ресурсам
- Намерение пользователя
- Область промпта

Однако, если источник различий не может быть объяснен или языки с ограниченными ресурсами систематически испытывают:

- неправильную идентичность,
- неправильное разрешение,
- недостаток материалов,
- высокую критическую ошибку

если это имеет место, существует проблема языковой справедливости. NOMOS определяет языковую справедливость следующим образом:

> *При равных материальных условиях сам язык не должен быть причиной систематического искажения.*

29. ПЕРЕСЕЧЕНИЕ ЯЗЫКА И СТРАНЫ

Языковой показатель не является независимым от страны. Один и тот же язык:

- разные страны,
- разный доступ к продукту,
- разные локальные ресурсы,
- разное законодательство,
- разные цены

может использоваться в контексте. Следовательно, ячейка «страна–язык»: (c,l) должна рассматриваться отдельно, когда это необходимо. Пример:

- Германия × турецкий
- Турция × турецкий
- Великобритания × турецкий

Это тот же язык, но может иметь различный локальный и рыночный контекст. Справедливость языка не может измеряться исключительно официальным языком страны.

30 наблюдатель страны И ПАНЕЛЬ СПРАВЕДЛИВОСТИ ЯЗЫКА ПЕРЕСЕЧЕНИЕ

Наблюдение одновременно:

- небольшое наблюдение Наблюдателя страны,
- наблюдение для языка с низким ресурсом

может иметь место. В этом случае запись должна содержать следующие поля:

PRIMARY_PANEL_ROLE

SECONDARY_DIAGNOSTIC_TAGS

вес населения Статус наблюдателя Статус языковой справедливости Одно и то же наблюдение не может быть полностью учтено в двух отдельных панелях. Это одно наблюдение. Оно имеет более одного диагностического тега.

31. ИЗМЕНЕНИЕ УЧАСТНИКОВ В ПАНЕЛИ НАБЛЮДАТЕЛЕЙ И ЯЗЫКА

Участник может быть изменен только из-за нарушения протокола. Следующие результаты не являются причинами для изменения:

- Неверный ответ ИИ
- Отказ
- Ответ на другом языке
- Критическая ошибка
- Путаница сущностей
- Отрицательная рекомендация

Это результаты измерений. Цель панели наблюдателей заключается именно в выявлении неожиданных событий. Изменение пользователя из-за того, что произошло неожиданное событие, подрывает цель панели.

32. ЯЗЫКОВАЯ ЭКСПЕРТИЗА

Экспертиза в Панели языковой справедливости должна как минимум соответствовать следующим условиям:

- Способность понимать промпт и ответ естественно или на высоком уровне владения

- Знание местной терминологии
- Способность оценивать источники доказательств
- Использование одного и того же нормативного руководства по оценке
- Быть слепым к модели и, если возможно, к интересам бренда
- Раскрытие конфликта интересов
- Путь третьего арбитра в случае разногласий

Язык, на котором невозможно найти арбитра:

ОТСУТСТВИЕ_ЯЗЫКОВОЙ_ЭКСПЕРТИЗЫ_ОЦЕНЩИКОВ

должен иметь статус. Нельзя предполагать, что все нюансы корректны, если перевод выполнен на английский автоматически. Машинный перевод может быть полезным инструментом. Он не является окончательной проверкой местного значения.

33. ПАРИТЕТ ИСТОЧНИКОВ

Объекта на английском сайте:

- объем охвата,
- цена,
- лицензия,
- ограничения,
- доказательства

существует; если они отсутствуют в версии на другом языке, низкий результат ИИ может быть не только проблемой модели. Требуются три отдельных результата:

- Сущность Язык Источник Разрыв
- ИИ-продукт Язык Представление Разрыв
- Общий Разрыв для Пользователя

Конечная проблема с точки зрения пользователя всё ещё реальна. Однако ответственность за исправление может варьироваться:

- сущность,
- Поставщик ИИ,

оба.

34. ПАРИТЕТ ИСКА

промпты, которые не являются семантически эквивалентными, дают разные ответы. Пример: на одном языке: «Чем занимается эта компания?» На другом языке: «Надёжна ли эта компания и стоит ли её рекомендовать?» Если спросить, результаты нельзя сравнивать. Конституция промптов будет подробно описана в следующем разделе. Положение этого раздела следующее:

> *Различия в производительности языковой модели не могут быть отнесены на счет производительности ИИ-языка без проверки эквивалентности промптов.*

35. ОТКЛОНЕНИЕ ЯЗЫКА ОТВЕТА

Пользователь даёт промпт на турецком языке. ИИ генерирует ответ на английском. Этот результат:

- даже если фактическая информация верна,
- может повлиять на опыт пользователя,
- доступность,
- и понимание.

Отклонение языка ответа должно быть зафиксировано отдельно:

СОВПАДЕНИЕ_ЯЗЫКА

ЧАСТИЧНОЕ_ИСПОЛЬЗОВАНИЕ_ДРУГОГО_КОДА

НЕЗАЯВЛЕННОЕ_ИСПОЛЬЗОВАНИЕ_ДРУГОГО_ЯЗЫКА

НЕПРИГОДНЫЙ_ЯЗЫКОВОЙ_ВЫХОД

НЕИЗВЕСТНО

Смена кода не всегда является ошибкой. Например, использование технических терминов на английском может быть нормальным. Это важно, если это нарушает основную удобопользуемость ответа.

36. ПИСЬМЕННАЯ СИСТЕМА И ТРАНСЛИТЕРАЦИЯ

Один и тот же язык может использоваться в разных системах письма. Для сущности:

- название,
- название продукта,
- юридическая идентичность

это может быть спутано с другой сущностью при транслитерации. Комитет по языковой справедливости, в той мере, в какой это применимо:

- система письма,
- транслитерация,
- официальное местное название
- имя, используемое пользователем

должен записывать свои поля. Различия в системах письма должны считаться только формальными. Если это влияет на разрешение сущности, это существенно.

37. SYNTHETIC Country Observer CYCLE

ПРЕЗЕНТАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Следующие продукт, страны, числа и результаты являются вымышленными. В реестре синтетического продукта под названием Orion AI Естественный охват: 120 допустимых стран. Основная Панель Населения зафиксировала допустимые наблюдения в первой волне только в 38 странах из-за взвешивания по населению.

37.1. Объем первой волны

Количество допустимых стран: 120 Стран с хотя бы одним наблюдением: 38

```
COC_count= 38/120= 31.7%
```

Пусть эти 38 страны составляют 97% от допустимого населения.

```
COC_pop= 97%
```

Результат: количество стран невелико, охват населения высок.

37.2. Четырехволновой цикл наблюдателей

Распределите оставшиеся 82 стран на четыре волны, используя ранее замороженную случайную и основанную на рисках программу. Каждая страна наблюдается как минимум один раз в течение цикла. В конце четвертой волны:

```
COC_count= 120/120= 100%
```

Однако достаточная выборка для оценки страны была достигнута только в 14 странах.

CEC = 14/120= 11.7%

Корректная публичная запись: покрытие наблюдения стран: 100%/ покрытие наблюдения населения: около 100%/ покрытие прогноза по странам: 11.7% Некорректная публичная запись: «Все 120 стран имеют надежный показатель страны».

37.3. Малый страновый критический сигнал

В записи Country Observer наблюдение AI в маленькой стране J ошибочно идентифицирует проверяемое лицо как лицензированную юридическую фирму. На самом деле, это лицо не предоставляет юридические услуги. Одного наблюдения недостаточно для оценки уровня ошибок в стране J; оно устанавливает кандидатное критическое событие. Протокол: проверить наблюдение; подтвердить промпт и справочную запись; повторить с новыми независимыми пользователями; если событие повторяется, открыть целевую выборку по стране; и оценить влияние на критическое соответствие отдельно. Наблюдаемое событие может быть редким; потенциальный вред не обязательно должен быть малым.

38. ДЕМОНСТРАЦИЯ СПРАВЕДЛИВОСТИ СИНТЕТИЧЕСКОГО ЯЗЫКА

СИНТЕТИЧЕСКАЯ ВЫБОРКА — НЕ РЕАЛЬНЫЕ РЕЗУЛЬТАТЫ ИИ ИЛИ КОМПАНИИ

Включите пять языков в тестирование Core Identity Synthetic Apple.com. Распределение главной панели населения:

Язык	Вес населения	Наблюдение на главной панели
L1	55%	550
L2	25%	250
L3	10%	100
L4	6%	60
L5	4%	40
Итого	100%	1,000

Результаты синтетического языка:

Язык	Доля успешных результатов
L1	95%
L2	92%
L3	88%
L4	65%
L5	58%

Глобальный результат с учетом численности населения:

```
0.55(95)+0.25(92)+0.10(88)+0.06(65)+0.04(58)=90.27
```

Глобальный показатель высок. Однако наблюдается сильная деградация в двух языках.

38.1. Пересборка данных для справедливости языков

Добавьте дополнительные наблюдения 140 к L4, 160 к L5. Новое количество наблюдений:

Язык	Всего наблюдений
L1	550
L2	250
L3	100
L4	200
L5	200
Итого	1,300

Грубое среднее значение по всем наблюдениям 1,300 ниже показанного выше глобального результата с учетом численности населения. Это не показатель глобального населения, так как L4 и L5 были намеренно пересобраны. После восстановления весов целевой популяции глобальный результат возвращается к указанной выше величине с учетом численности населения. Пересборка повышает чувствительность для L4 и L5 без изменения оценки с учетом численности населения.

38.2. Таблица оценки языковой справедливости

Синтетический результат. Глобальный показатель с учетом численности населения: 90.27. Показатель равного языка:

```
(95+92+88+65+58)/5=79.6
```

Разница между языками от самого высокого к самому низкому: 37 очков; язык с наименьшим показателем: 58; среднее значение для двух нижних языков: 61.5; уровень критических ошибок в языках с ограниченными ресурсами: указывается отдельно; эквивалентность промпта: предполагается подтвержденной; паритет источников: предполагается подтвержденным. Правильная интерпретация:

> *Представленность мирового населения кажется высокой; однако справедливость по языкам крайне слабая.*

Неверная интерпретация: «Поскольку глобальный показатель равен 90, GEO успешен на всех языках.»

39. СИНТЕТИЧЕСКИЙ СЛУЧАЙ, ВЫЯВЛЯЮЩИЙ ИСТОЧНИК УХУДШЕНИЯ ЯЗЫКА

Рассмотрим три языка.

Язык A

Эквивалент промпта Официальный источник полный Местный арбитр достаточен Ответ ИИ неверен Результат: кандидат на проблему представления языка в продукте ИИ.

Язык B

Перевод запроса существенно отличается от ответа ИИ, соответствующего запросу. Результат: недостаток в структуре запроса.

Язык C

На официальной языковой странице старая цена ИИ переносит старую цену Английский источник является актуальным Результат: проблема языка источника сущности и совокупный разрыв в представлении достигают пользователя. Даже если показатели трех языков кажутся одинаковыми, владелец исправления отличается.

40. ПАНЕЛИ СТРАН И ЯЗЫКОВ ОБЩЕДОСТУПНАЯ КАРТА РЕЗУЛЬТАТОВ

В публичной карточке результата кандидата должны быть указаны следующие поля:

Страновые поля

Естественный охват соответствующее количество стран Количество стран с как минимум одним наблюдением Покрытие наблюдений по странам Покрытие наблюдений с учетом численности населения Покрытие прогноза по странам Цикл времени наблюдателя Статус уведомлений наблюдателя ЗА ПРЕДЕЛАМИ РОДНОГО ДОСТУПА страны, НЕ НАБЛЮДАЕМЫЕ страны Дата измерения

Языковые поля

Количество охваченных языков Наблюдаемое количество языков Количество языков, для которых можно сделать прогнозы Официально поддерживаемые языки Экспериментальные языки Результат по мировому языку с учетом численности населения Результат при равном признании языков Минимальное количество языков Различие между языками Критические показатели ошибок

ОТСУТСТВИЕ_ЯЗЫКОВОЙ_ЭТАЛОННОЙ_БАЗЫ

ОШИБКА_ЭКВИВАЛЕНТНОСТИ_ПОДСКАЗОК

НЕТЕСТИРОВАННЫЕ языки Версия панели языков Эта карта делает видимой реальность, которую не может нести единственный «многоязычный» значок.

41. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH07-N01

Панели Country Observer и Language Fairness должны быть определены отдельно от основной панели населения с точки зрения цели, выборки и весов.

CH07-N02

Наблюдение странового наблюдателя не может использоваться в качестве оценки населения страны.

CH07-N03

Для страны наличие наблюдения и способность составлять оценку по стране — это разные статусы.

CH07-N04

Количество охваченных стран и взвешенное по населению покрытие стран должно указываться вместе.

CH07-N05

Оценка охвата по стране должна отображаться отдельно от охвата наблюдений.

CH07-N06

Каждая соответствующая критериям страна должна иметь вероятность быть наблюдаемой больше нуля в определённом цикле наблюдения; страны вне зоны охвата также должны быть показаны.

CH07-N07

Цикл наблюдения и классы частоты по странам должны быть версионированы перед просмотром ответов ИИ.

CH07-N08

Страна с низким баллом не может быть удалена из цикла Наблюдателя после того, как результаты будут увидены.

CH07-N09

Оповещение Наблюдателя страны не является автоматической неудачей страны; оно запускает проверку и эскалацию.

CH07-N10

Одно событие Critical Observer также может инициировать отдельную проверку соответствия без оценки распространенности.

CH07-N11

Наблюдения с Панели наблюдателей страны не могут использоваться в основной Популяционной панели с равным исходным взвешиванием.

CH07-N12

Если наблюдения странового наблюдателя включены в основной балл, их вероятности выбора и веса должны быть явно смоделированы.

CH07-N13

Справедливость в языке нельзя определять как обязанность давать один и тот же ответ слово в слово.

CH07-N14

Справедливость языка требует сопоставимой точности, ограничений, источника и пригодности для пользователя при эквивалентных материальных условиях.

CH07-N15

Язык и страна, язык и локаль, а также язык и система письма не могут использоваться взаимозаменяемо.

CH07-N16

Вселенная языков, охватываемая измерением, и причина включения каждого языка должны быть объяснены до проведения измерения.

CH07-N17

Утверждение «все языки» может использоваться только в том случае, если действительно измерялись все соответствующие языковые и пользовательские интерфейсы; в противном случае следует говорить «языки в рамках области охвата».

CH07-N18

Наблюдения на языке, который официально не поддерживается, не могут быть добавлены к основной производительности поддерживаемого продукта без объяснения.

CH07-N19

Чрезмерная выборка языков с ограниченными ресурсами не должна увеличивать глобальный вес населения.

CH07-N20

Если наблюдения с панели справедливости языка включены в глобальный результат, они должны быть возвращены к фактическим весам населения целевого языка.

CH07-N21

Один и тот же многоязычный пользователь не может учитываться более чем за одного полного человека в глобальном весе населения.

CH07-N22

Без проверки эквивалентности промптов разницу в языковом балле нельзя однозначно отнести к производительности языкового ИИ.

CH07-N23

Низкий языковой балл нельзя возлагать только на продукт ИИ без оценки справедливости по эталону.

CH07-N24

Решение о точности языка не может быть финализировано без достаточной компетенции местного арбитра.

CH07-N25

Отклонение в языке ответа и переключение кода должны регистрироваться отдельно.

CH07-N26

Законные различия в местной действительности не могут классифицироваться как языковая несправедливость.

CH07-N27

Необъяснимые существенные языковые различия нельзя списывать исключительно на персонализацию или различия стиля.

CH07-N28

Глобальный балл населения не может заменять результат языковой справедливости.

CH07-N29

Равный языковой балл нельзя представлять как глобальный балл с учетом веса населения.

CH07-N30

Необходимо показывать показатели критических ошибок на основе языка отдельно от общего среднего значения.

CH07-N31

PROMPT_EQUIVALENCE_FAILURE, REFERENCE_LANGUAGE_GAP, MEASUREMENT_PARITY_FAILURE и ADJUDICATION_LANGUAGE_GAP должны поддерживаться как отдельные статусы.

CH07-N32

В панелях Country Observer и Language Fairness решение о смене участников не может основываться на правильности ответа ИИ или на пользе для бренда.

CH07-N33

Одна и та же наблюдательная запись может иметь несколько диагностических меток; однако она не может получать больше одного полного веса в глобальном прогнозе без объяснения.

CH07-N34

Записи панелей страны и языка должны иметь ответственное человеческое или институциональное лицо.

42. ФОРМЫ ОШИБОК

CH07-F01— ОЦЕНКА СТРАНЫ ОДНИМ ЧЕЛОВЕКОМ

Наблюдение одного наблюдателя публикуется как оценка GEO для страны.

CH07-F02— СИМВОЛИЧНОЕ ПОКРЫТИЕ СТРАНЫ

Каждому человеку назначается одна страна, и это называется «пропорционально населению мира».

CH07-F03— ПУТАЮТ ПОКРЫТИЕ НАСЕЛЕНИЯ С КОЛИЧЕСТВОМ СТРАН

Хотя наблюдается небольшое количество крупных стран, показано только покрытие населения; хотя наблюдается большое количество малых стран, показано только количество стран.

CH07-F04— ПРИСВАИВАНИЕ ОБЛАСТИ НАБЛЮДЕНИЯ КАК ОБЛАСТИ ОЦЕНКИ

Все страны с хотя бы одним наблюдением считаются оцененными странами.

CH07-F05— УЧЁТ НАБЛЮДЕНИЯ ОТ НАБЛЮДАТЕЛЯ СТРАНЫ ДВАЖДЫ В ОСНОВНОЙ ОЦЕНКЕ

Одно и то же наблюдение получает полный вес как в панели наблюдателей, так и в панели населения.

CH07-F06— ПРЕОБРАЗОВАНИЕ ПРЕДУПРЕЖДЕНИЯ ОТ НАБЛЮДАТЕЛЯ СТРАНЫ В УТВЕРЖДЕНИЕ О РАСПРОСТРАНЁННОСТИ

Одно событие обобщается на всех пользователей страны.

CH07-F07— ПРОИГНОРИРОВАТЬ СОБЫТИЕ КРИТИЧЕСКОГО НАБЛЮДАТЕЛЯ

Высокий риск вреда не рассматривается при утверждении «только один пользователь».

CH07-F08— УДАЛЕНИЕ СТРАНЫ С НИЗКИМ БАЛЛОМ ИЗ ЦИКЛА

График наблюдателя изменяется в зависимости от результатов.

CH07-F09— НАБЛЮДЕНИЕ ТОЛЬКО ЗА КОММЕРЧЕСКИМИ РЫНКАМИ

Страны, имеющие доступ к продуктам, но считающиеся коммерчески неважными, полностью исключаются; результат представляется глобально.

CH07-F10— РАССМАТРИВАНИЕ СТРАНЫ И ЯЗЫКА КАК ОДНОГО И ТОГО ЖЕ

Официальный язык страны принят как язык всех пользователей.

CH07-F11— УТВЕРЖДЕНИЕ ВСЕХ ЯЗЫКОВ

Измеряются только некоторые ограниченные языки; результат публикуется как «на всех языках».

CH07-F12— УДАЛЕНИЕ НИЗКОГО СВАРНОГО ЯЗЫКА

Поскольку выборка мала, язык полностью исключен из отчета.

CH07-F13— ПЕРЕРАСТАЮЩИЙ ЯЗЫК С НЕДОСТАТОЧНЫМИ РЕСУРСАМИ В МИРОВОМ МАСШТАБЕ

Язык с избыточной выборкой учитывается в глобальном индексе в соответствии с его долей в выборке.

CH07-F14— УЧЕТ РАВНОГО ЯЗЫКОВОГО ИНДЕКСА КАК ИНДЕКСА НАСЕЛЕНИЯ

Диагностический индекс, который придает равный вес каждому языку, представлен как средний опыт мировой популяции.

CH07-F15— УЧЕТ ИНДЕКСА НАСЕЛЕНИЯ КАК СПРАВЕДЛИВОСТЬ ЯЗЫКА

Глобальный результат, усиленный весом крупных языков, скрывает искажение для малых языков.

CH07-F16— УЧЕТ ОШИБКИ ПЕРЕВОДА промпта КАК ОШИБКИ ИИ

Разница, возникающая из-за семантически различных промптов, приписывается модели.

CH07-F17— УЧЕТ РАЗРЫВА ИСХОДНОГО ЯЗЫКА КАК ОШИБКИ ИИ

Хотя локализованной официальной информации нет, вся ответственность возлагается на продукт ИИ.

CH07-F18— ЗАЩИТА ОШИБКИ ИИ ИЗ-ЗА НЕДОСТАТКА РЕСУРСОВ

Несмотря на наличие тех же и достаточных ресурсов, ошибки на языке с низким уровнем ресурсов закрываются исключительно как отсутствие контента.

CH07-F19— РАССМОТРЕНИЕ НЕАДЕКВАТНОСТИ ЯЗЫКА АДЖЮДИКАТОРА КАК СБОЯ ПРОДУКТА

Даже если аджюдикатор неправильно понял, результат считается окончательным.

CH07-F20— РАССМОТРЕНИЕ МАШИННОГО ПЕРЕВОДА КАК ОКОНЧАТЕЛЬНОГО СУДЬИ

Потеря локального значения не проверяется.

CH07-F21— РАССМОТРЕНИЕ ИЗМЕНЕНИЙ В КОДЕ КАК АВТОМАТИЧЕСКОЙ ОШИБКИ

Сбои происходят при естественных технических терминах или допустимом смешении языков.

CH07-F22— РАССМОТРЕНИЕ НЕПРИГОДНОГО ЯЗЫКОВОГО ОТКЛОНЕНИЯ КАК СТИЛЯ

Даже если система отвечает на другом языке, который пользователь не может понять, результат считается успешным.

CH07-F23— ИГНОРИРОВАНИЕ РАЗЛИЧИЯ В ПИСЬМЕННОЙ СИСТЕМЕ

Ошибка транслитерации, вызывающая путаницу сущностей, считается формальной.

CH07-F24— УЧЁТ МНОГОЯЗЫЧНОГО ПОЛЬЗОВАТЕЛЯ ДВАЖДЫ

Один и тот же человек учитывается как отдельная единица населения на всех языках.

CH07-F25— РАССМОТРЕНИЕ ЗАКОННОГО ЛОКАЛЬНОГО РАЗЛИЧИЯ КАК НЕСПРАВЕДЛИВОГО

Даже если разница в цене или законе основана на реальных местных условиях, это объявляется языковой предвзятостью.

CH07-F26— РАССМОТРЕНИЕ НЕОБЪЯСНИМОГО РАЗЛИЧИЯ КАК ЛИЧНОГО

Материальная несогласованность при равных условиях не ставится под сомнение.

CH07-F27— ОПРЕДЕЛЁННЫЙ БАЛЛ ИЗ-ЗА НЕХВАТКИ ЯЗЫКОВОЙ ЯЧЕЙКИ

Рейтинг языков проводится на основе нескольких наблюдений.

CH07-F28— НАСТРОЙКА ЯЗЫКОВОЙ ПАНЕЛИ ПОСЛЕ РЕЗУЛЬТАТА

Выбираются только языки с низким или высоким числом баллов.

CH07-F29— ИЗМЕРЕНИЕ ТОЛЬКО ЯЗЫКОВ С ВЫСОКИМ ИСТОЧНИКОМ

Результат представлен как «многоязычный».

CH07-F30— СОХРАНЕНИЕ НЕПРОВЕРЕННЫХ ЯЗЫКОВ

Поле НЕ ТЕСТИРОВАНО не отображается на публичной карточке результата.

43. ПРОЦЕДУРА АУДИТА

Шаг 1— Заморозка реестра соответствия продукта и страны

Для каждой страны:

- Естественный охват
- доступ к продукту
- язык аудита
- пользовательский интерфейс

статус записан.

Шаг 2— Определение охвата стран основной панели населения

Определяется, в каких странах наблюдается основная панель.

Шаг 3— Определение универсума стран наблюдателей

Определены страны, которые имеют недостатки в основной панели и имеют статус CE-3.

Шаг 4— Установить цикл наблюдателя

Страны:

- назначаются по частотному классу,
- волне,
- языку/локали,
- пользовательский интерфейс

и поддержки.

Шаг 5— Предварительный выбор участников-наблюдателей

Основные и резервные списки создаются до просмотра результатов.

Шаг 6— Классификация результатов наблюдателей

Назначаются доступ, идентификация, язык, материальный факт, статусы критических событий.

Шаг 7— Применение сигнала тревоги и эскалации

По типу сигнала тревоги:

- снова,
- целевой образец
- контроль справочных материалов,
- проверка соответствия

запускается.

Шаг 8— Вычислить показатели охвата стран

Объем количества стран, объем населения и объем оценки страны рассчитываются отдельно.

Шаг 9— Заморозить запись языка

Для каждого языка:

- по причинам охвата,
- статус поддержки,
- система письма,
- локаль,
- пользовательская совокупность

записываются.

Шаг 10— Установить дополнительную выборку для справедливости языка

Дополнительные цели устанавливаются для языков, недостаточно представленных в основной панели населения.

Шаг 11— Проверить четыре языковых критерия

Оцениваются системная эквивалентность, референтная паритетность, паритетность измерения и паритетность судебного решения.

Шаг 12— Оценка языковых наблюдений

Для каждого языка:

- правдивость,
- недостаток материалов,

- Критическая ошибка,
- неизвестно,
- отклонение языка ответа

записываются.

Шаг 13— Разделить глобальные и диагностические веса

Результат, взвешенный по численности населения, и результат с равным признанием языков рассчитываются отдельно.

Шаг 14— Рассчитать языковое неравенство

языковой порог, значительная языковая разница, различие критических ошибок, рассчитывается коэффициент неизвестного.

Шаг 15— Исследовать пересечения страна–язык

Определяются различия или конфликты в локализации материалов.

Шаг 16— Создать публичную карточку результата

Область применения, достаточность прогноза, сигналы тревоги и не протестированные области становятся видимыми.

44. НЕОБХОДИМЫЕ ДОКАЗАТЕЛЬСТВА

Запись продукта ИИ и пользовательского интерфейса; регистрация страны Естественный охват; количество правомочных стран; статус доступа к стране; распределение основных стран Панели населения; наблюдаемая совокупность страны; цикл наблюдателя; классы частоты наблюдателей; матрица распределения стран по волнам; распределение стран по языкам; основные и резервные списки наблюдателей; результаты наблюдений по странам; оповещения наблюдателей; записи об эскалации; целевые выборки по странам; охват наблюдений по странам; охват стран с учетом населения; охват стран для оценок; реестр языков в рамках охвата; статус поддержки языков; записи о скриптах и локали; основные подсчеты выборки по языкам; подсчеты превышения выборки для справедливости языка; веса населения по языкам; записи многоязычных пользователей; записи эквивалентности промптов; записи справедливости по референсу; записи справедливости измерений; записи квалификации местных оценщиков; и результаты по языкам.

Языковые критические показатели ошибок Отклики на языке отклонения Языковое неравенство Метрики Неопробованные языки Равные результаты по языку и населению Веса панели Проверки на двойной учет Публичная карточка результатов Версии панели Ответственное лицо или учреждение

45. ЧЕК-ЛИСТ АУДИТА

Была ли цель Панели наблюдателей по странам отделена от основной панели населения? Какие страны включены в Панель наблюдателей? Актуален ли статус права на участие и доступ для стран? Существует ли вероятность положительного наблюдения в каждом цикле наблюдения соответствующей страны? Была ли остановка цикла наблюдения до получения результатов? Использовалось ли наблюдение Панели наблюдателей по странам как оценка страны? Были ли разделены охват наблюдений стран и охват оценок страны? Был ли охват

стран, взвешенный по населению, показан отдельно? Были ли наблюдения Панели наблюдателей по странам учтены дважды в основной оценке? Было ли подтверждено критическое предупреждение наблюдателя? Был ли один сигнал тревоги преобразован в утверждение о распространенности? Было ли открыто целевое выборочное обследование из-за сигнала тревоги? Был ли заранее определен круг изучаемых языков? Соответствует ли утверждение «все языки» фактическому охвату? Были ли разделены язык, письменная система и локаль?

Был ли зафиксирован статус поддержки языков продукта? Почему были выбраны языки с ограниченными ресурсами? Были ли примеры языковой справедливости перераспределены в глобальном рейтинге? Был ли один и тот же многоязычный пользователь посчитан дважды? Была ли проверена эквивалентность промптов? Соответствуют ли официальные источники по содержанию на разных языках? Знают ли администраторы соответствующий язык на достаточном уровне? Использовался ли машинный перевод в качестве окончательного судьи? Были ли зафиксированы отклонения в языке ответа? Было ли легитимное местное разнообразие отделено от языковой несправедливости? Были ли опубликованы языковой корпус и вариативность языка? Видны ли показатели критических ошибок на основе языка? Были ли разделены REFERENCE_LANGUAGE_GAP и другие ошибки измерения? Видны ли нетестированные языки? Скрывает ли публичный рейтинг проблему языковой справедливости?

Получило ли одно и то же наблюдение полный вес в нескольких панелях? Идентифицирован ли ответственный владелец записей панели?

46. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Разве дать одного человека каждой маленькой стране не лучше, чем не давать никому?»

Для раннего наблюдения и географического охвата — да. Но этот человек не измеряет долю страны. Правильный термин:

Наблюдение Country Observer

должен быть.

Возражение 2— «Если один человек не дает оценку страны, зачем мы его отправляем?»

Потому что одно наблюдение может:

- выявить проблемы с доступом,
- прояснить путаницу с объектами,
- опознать неправильный язык,
- и зафиксировать критические события

видно впервые. Панель Наблюдателя является системой раннего предупреждения.

Возражение 3— «Должны ли мы наблюдать за каждой страной в каждой волне?»

Нет. Можно использовать определённый цикл наблюдения. Страны с высоким риском или значимые страны можно наблюдать чаще; маленькие и низкорисковые страны можно наблюдать поочередно. Расписание должно быть определено до получения результатов.

Возражение 4— «Если маленькие страны имеют низкий вес в глобальном рейтинге, почему мы уделяем им так много внимания?»

Потому что их доля населения может быть низкой. Однако:

- закон,
- языка,
- местный ущерб для пользователей,
- географическая целостность

являются реальными человеческими сообществами. Низкий вес не должен означать невидимость.

Возражение 5— «Разве одинаковое взвешивание всех языков не является самым справедливым методом?»

Равное взвешивание языков может быть ценным диагностическим подходом. Оно не отражает средний человеческий опыт среди мирового населения. Оценка для всего населения и оценка языковой справедливости должны публиковаться отдельно.

Возражение 6— «Если небольшой язык лишь немного влияет на глобальный результат, почему низкий балл имеет значение?»

Потому что проблема представления реальна для людей, использующих этот язык. Кроме того, серьёзные ошибки в лицензиях, здравоохранении или идентификации не становятся автоматически незначительными из-за доли населения.

Возражение 7— «Разве несправедливо обвинять ИИ, если язык с низким ресурсом имеет мало ресурсов?»

Обвинять ИИ может быть несправедливо. Поэтому:

- паритет ресурсов,
- эквивалентность промптов,
- компетентность арбитра

следует изучать отдельно. Ошибка, которую видит пользователь, остаётся фактически правильной. Владелец исправления должен быть определён корректно.

Возражение 8— «Разве нормально получать разные ответы на одном и том же языке в разных странах?»

Это может быть нормально, если реальные местные условия различаются. Если один и тот же материальный факт меняется по-разному и без объяснения, это может быть проблемой. Язык и локаль должны рассматриваться вместе.

Возражение 9— «Если ИИ дает правильный ответ на другом языке, почему он должен ошибаться?»

Если пользователь не может понять ответ, представление не может быть использовано. Отклонение языка следует оценивать отдельно вместе с точностью.

Возражение 10— «Поиск пользователей 200 для языковых оценок может быть очень дорогим.»

Можно использовать меньшую выборку. Результат:

- обнаружение,
- временный,
- недостаточная выборка

должен публиковаться с этим статусом. Стоимость не дает ложной уверенности.

Возражение 11— «Если глобальный результат уже высок, не пугает ли отдельная языковая карта клиента без необходимости?»

Сокрытие ухудшения материала не защищает клиента. Это создает ложное чувство уверенности. Можно использовать короткую и ясную карту результатов.

Возражение 12— «Сведет ли одно событие Критического Наблюдателя ко всем критериям соответствия?»

Автоматический ответ зависит от типа события. Одно событие не указывает на национальное распространение. Однако если оно влечет за собой серьезный ущерб или ложные полномочия, это может инициировать критический пересмотр и временные меры. Решение будет определено в разделе 15.

ОБЩЕЕ ПРАВИЛО РАЗДЕЛА 49

Создание мирового рейтинга легко. Вы заполняете мировое население большими странами. Вы теряете малые страны по частям. Вы оцениваете языки с большим количеством ресурсов. Языки с низким количеством ресурсов: вы убираете, говоря «Недостаточно данных». Затем вы публикуете одно число. Это число может быть математически рассчитано правильно. Но оно может не показывать всех недостатков представления мира. NOMOS не придаёт равного веса каждой стране. Потому что доли населения людей не равны. NOMOS также не игнорирует малые страны. Потому что низкое население не означает несуществование. NOMOS не учитывает все языки одинаково в мировом рейтинге. Потому что это может исказить средний опыт мирового населения. NOMOS также не делает малые языки невидимыми. Потому что даже если их глобальный вес низок, ошибка представления для людей, использующих этот язык, может быть 100% реальной.

Панель наблюдателей страны:

> *Показывает там, где мы никогда не смотрели.*

Панель языковой справедливости:

> *Показывает, кто находится ниже глобального среднего.*

Панель населения:

> *Оценивает средний опыт в мировой популяции.*

Когда эти три работают вместе, глобальный показатель становится одновременно достоверным для населения и проверяемым с точки зрения справедливости. Один пользователь в стране — это не вся страна. Но этот пользователь — реальный человек. Язык может составлять небольшую часть глобального населения. Но ошибка в этом языке не становится менее неправильной. Критическое событие может быть редким. Но вред измеряется не только частотой. Поэтому седьмой закон измерений NOMOS заключается в следующем:

> *Право на видимость не то же самое, что вес в популяции.*

Его восьмой закон таков:

> *Наблюдение за страной не то же самое, что оценка страны.*

Девятый закон звучит следующим образом:

> *Языковая справедливость — это не равенство всех языков; это обеспечение того, чтобы никто систематически не отставал от материальной реальности из-за своего языка.*

Десятый закон таков:

> *Если глобальное среднее делает язык и страну, которые отстают, невидимыми, этого самого по себе недостаточно для географического результата.*

Раздел NOMOS Заказ 7

> *Не скрывайте страны, на которые вы никогда не смотрите, показывая мне мировой рейтинг.*
>
> *Не отправляйте одного пользователя в маленькую страну и не делайте этого человека представителем всей страны.*
>
> *Но также не игнорируйте маленькую страну навсегда только потому, что её население мало.*
>
> *Разделяйте наблюдение за страной, прогноз по стране и соответствие страны.*
>
> *Не судите всю страну по одной серьёзной ошибке. / Но и не игнорируйте вред людям, говоря «только один ответ».*
>
> *Не основывайте успех в крупных языках на молчании малых языков.*
>
> *Вы можете много выбирать для языка с низкими ресурсами. / Но не делайте так, чтобы он казался более крупным, чем есть на самом деле, в мировом населении.*
>
> *Не превращайте одинаковую языковую оценку в оценку населения; не превращайте оценку населения в оценку языковой справедливости.*
>
> *Не вините меня, если перевод неправильный. / Если источник неполный, также укажите это. / Если арбитр не понимает язык, не делайте окончательного суждения.*
>
> *Не классифицируйте законное местное различие как несправедливость, и не списывайте необъяснимое значительное расхождение как простую стилистику.*
>
> *Сначала наблюдайте за странами. / Затем укажите, где вы можете поддержать оценку. / Затем сделайте сверхвыборку языков, требующих диагностической видимости. / Затем верните их к истинным весам населения. / Держите на виду людей, которых средний мировой результат оставляет в тени.*

Заключительное предложение главы

По-настоящему глобальное GEO-измерение должно показывать больше, чем видит большинство мира. Оно также должно показывать, кто остаётся незамеченным в меньших странах и языках, кто представлен неверно и какой результат является лишь единичным предупреждением.

Нормативное ядро

Панели населения GEO-1000, Панели наблюдателей по странам и Панели языковой справедливости ДОЛЖНЫ оставаться различными по цели, выборке, взвешиванию и публичным утверждениям. Наблюдение Панели наблюдателей по странам МОЖЕТ установить, что условие доступа, ответ, аномалия или критический инцидент произошли в определенном контексте страна–продукт–язык. Наблюдение Панели наблюдателей по странам НЕ ДОЛЖНО представляться как оценка распространенности по стране или как показатель GEO страны. Покрытие наблюдений по странам, покрытие наблюдений с учетом веса населения и покрытие оценки по стране ДОЛЖНЫ быть представлены отдельно. Каждая подходящая страна ДОЛЖНА получать ненулевую вероятность включения в наблюдение в рамках заранее объявленного цикла наблюдения, если она не выходит за пределы правомерного естественного охвата продукта. Языковая справедливость ДОЛЖНА оценивать материально сопоставимое представление в рамках семантически эквивалентных запросов и сопоставимых контекстов. Она НЕ ДОЛЖНА требовать точно идентичных ответов слово в слово или игнорировать законные местные различия. Языки с избыточной выборкой ДОЛЖНЫ быть возвращены к их целевым весам на популяцию для глобальных оценок. Оценки с учетом веса популяции, диагностика при равных весах языков, минимальные уровни по языкам, показатели критических ошибок, показатели неизвестных элементов и непроверенных языков ДОЛЖНЫ оставаться отдельно видимыми. Различие в языке НЕ ДОЛЖНО приписываться исключительно продукту ИИ до тех пор, пока не будут оценены эквивалентность запросов, паритет эталонных данных, паритет измерений и паритет рассмотрения. Одно наблюдение МОЖЕТ иметь несколько тегов диагностической панели, но НЕ ДОЛЖНО получать дублирующий вес всей популяции. Инциденты критического наблюдателя МОГУТ вызвать подтверждающую выборку и проверку соответствия без того, чтобы рассматривать их как доказательство распространенности в стране.

ГЛАВА 08

Отбор и взвешивание участников и смещение выборки

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K04; K05; K06; K07; K08; K22

Граница главы

Глава 5 определила целевую пользовательскую аудиторию. Глава 6 преобразовала эту аудиторию в выборочный план, распределённый по странам, языкам и пользовательским интерфейсам. Глава 7 установила дополнительные системы наблюдения для малых стран и языков с ограниченными ресурсами, которые основная панель населения могла бы оставить незамеченными. Теперь числа в плане выборки должны превратиться в реальных людей:

> *Где будут найдены эти участники, как они будут отобраны, на каких условиях их допустят в панель и с какими искажениями они войдут в GEO-оценку?*

Из онлайн-панели исследовательской организации? Через открытый призыв в социальных сетях? Из университетов? Из собственного списка пользователей клиента? От местных исследовательских партнёров? Только от людей, которые часто пользуются продуктами ИИ? От тех, кто использует продукт впервые? От клиентов бренда? От сотрудников конкурентов? От подписчиков платного плана? Как будет проверяться, действительно ли человек живёт в этой стране? Сколько раз один и тот же человек будет учитываться, если у него несколько аккаунтов? Сколько будет оплачено участнику? Будет ли оплата стимулировать быстрое выполнение задания или предоставление конкретного результата? Будет ли наблюдение сохранено, если участник загрузит неправильный ответ ИИ? Или его заменят новым пользователем, сказав: «Задача выполнена неправильно»? Могло быть собрано тысяча действительных скриншотов. Но участники:

- 60%— активные пользователи ИИ,
- 20% являются студентами технологического факультета,
- 10% являются клиентами проверяемого бренда,
- 5% являются сотрудниками компании,
- остальные — обычные пользователи

Какое население, тогда, будет представлять результат? Может ли демографическое взвешивание полностью исправить разницу? Сопоставление возрастных, гендерных, страновых и языковых показателей не обязательно исправляет немеряемые различия в интересе к технологиям, осведомленности о брендах, доступе к платным планам или схемах использования ИИ. Взвешивание ценно, но оно не может устранить все дефекты в выборе участников. Эта глава применяет ранее указанное требование — чтобы каждый контроль

аудита можно было проверить по модели, дате, стране, языку, набору запросов, количеству повторений и записи измерений — к выбору фактических участников. Она также сохраняет центральный требователь NOMOS: доказательства, границы, контекст и время. Эта глава:

- источники отбора участников
- различие между вероятностными и не-вероятностными панелями,
- проверка права на участие и удостоверение личности
- контроль доступа по стране, языку и продукту,
- уникальность участников,
- вознаграждение и риск перенаправления,
- кондиционирование панели,
- назначение заданий,
- недоответ и уход с панели,
- мошенническое или автоматизированное участие,
- смещение при выборе и покрытии,
- влияние и ограничения взвешивания,
- статусы качества выборки

определяет. В этой главе пока не рассматривается:

- все технические области реестра ИИ-продуктов,
- полная система перевода и версий промптов,
- подробная архитектура инфраструктуры NOMOS Capture,
- атомарное рассмотрение утверждений,
- формула окончательного балла

не полностью определена. Ключевой вопрос раздела 8 заключается в:

> *Когда мы собираем тысячу действительных записей, как мы доказываем, кому из реальных людей принадлежат эти записи и какую человеческую популяцию они могут представлять?*

NOMOS Вызов

Вы покупаете 1,000 человек у онлайн-панели для исследований. Провайдер панели сообщает вам: «1,000 участников готовы согласно вашим квотам по стране и возрасту». Участники выполняют задание. Они загружают скриншоты. Исходные результаты:

> *90 процентов точного представления.*

В вашем отчете вы пишете: «90% пользователей по всему миру точно представлены». Однако фактическое распределение участников следующее:

Частота использования ИИ	Доля целевой популяции	Доля в панели
Редкое или низкое использование	50%	10%
Среднее использование	30%	30%
Интенсивное использование	20%	60%

Активные пользователи ИИ:

- лучшие привычки к формулировке запросов,
- больше оплаченных аккаунтов,

- новые интерфейсы продукта,
- знания о персонализации,
- большая знакомость с технологиями

может быть. Даже если используется тот же зафиксированный запрос:

- план аккаунта,
- интерфейс,
- предыдущие использования,
- доступ к продукту

могут быть разными. Пусть пропуски синтетической группы будут следующими:

Частота использования ИИ	Доля успешных результатов
Редкое или низкое использование	72%
Среднее использование	84%
Интенсивное использование	96%

Исходный результат панели:

```
0.10(72)+0.30(84)+0.60(96)=90
```

то есть:

90 процентов

С учётом весов целевой популяции, однако, результат следующий:

```
0.50(72)+0.30(84)+0.20(96)=80.4
```

Те же ответы 1,000, следовательно, дают два разных результата:

- Исходный результат панели: 90 процентов
- Результат целевой популяции: 80.4 процентов

Теперь представьте, что вы применили взвешивание. Каждый из 100 человек в группе с низким использованием получает высокий вес. Каждый из 600 человек в группе с высоким использованием получает низкий вес. Исходный размер выборки по-прежнему составляет 1,000. Однако эффективный размер выборки снижается примерно до 349. Таким образом, вы могли собрать 1,000 файлов. Статистически, у вас может не быть объема информации, эквивалентного 1,000 равным и независимым людям. Теперь более сложный вопрос: в группе с низким использованием люди, которые никогда не были включены в панель:

- слабо подключенные,
- использующие вспомогательные технологии,
- живущие в маленьком городе,
- не имеющие конкретного устройства

Если есть люди, может ли взвешивание вернуть их? Нет. Наблюдение за человеком, который вообще не находится в выборочной совокупности, нельзя получить с помощью взвешивания. Теперь аудиторская компания отправляет приглашения своим клиентам: «Пожалуйста, участвуйте в опросе». Лояльные клиенты компании участвуют в опросе с высокой частотой. Люди, которые знают конкурентов, но не пользуются услугами компании, участвуют реже. Веса по возрасту и стране могут быть корректными. Смещение, связанное с лояльностью к бренду, может сохраняться. Теперь другая проблема: участнику говорят: «Вы получите бонус, если загрузите правильный и качественный ответ ИИ». Участник исправляет ответ. Он выбирает лучший результат. Он обрезает скриншот. Это больше не является естественной

реакцией продукта ИИ. Это выбранная реакция, созданная системой вознаграждений. Первое правило этого раздела:

> *Количество участников не является представительной силой участников.*

Второе положение гласит:

> *Случайное назначение не является случайным выбором популяции.*

Её третье положение гласит:

> *Взвешивание может уменьшить наблюдаемый дисбаланс; оно не может создать ненаблюдаемого человека.*

Четвертая статья гласит:

> *Включение участника в панель, его удержание в задании или замена не могут зависеть от содержания ответа ИИ.*

Пятая его статья гласит:

> *Награда должна выдаваться за честное выполнение протокола, а не за выбор положительного или желаемого ответа.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — сделать всю цепочку отбора от целевой популяции до фактического наблюдения за пользователем видимой, беспристрастной и проверяемой. В разделе нормативно различается:

- Целевая популяция и рамки выборки
- Рамка выборки против источника участников
- Выбранное лицо против приглашенного лица
- Приглашенное лицо и лицо, которое соглашается участвовать
- Участник и человек, выполняющий задачу
- Человек, который выполняет задачу, и действительное наблюдение
- Человек и аккаунт
- Человек и устройство
- Человек и сессия
- Вероятностная панель и панель добровольцев
- Случайный отбор внутри панели и случайный отбор из популяции
- Соблюдение квоты и истинное представление
- Демографический баланс и поведенческий баланс
- Личное знакомство и пригодность для целевой популяции
- Клиент, сотрудник и обычный пользователь
- Вознаграждение участника и стимул к результату
- Скорость выполнения задания и целостность задания
- Недействительное наблюдение с отсутствием ответа
- Изменение популяции с потерей панели
- Фальшивый участник и реальное наблюдение низкого качества
- Мошенничество с данными и неправильный ответ ИИ

- Коррекция смещения с помощью взвешивания
- Производство доказательств с калибровкой
- Сырая выборка против эффективной выборки
- Конфиденциальность участников и аудиторские доказательства
- Панель общей популяции и панель экспертов или клиентов

В конце этого раздела каждая проверка GEO-1000 должна быть способна ответить на следующие вопросы:

> *Из какого источника и по какому правилу были выбраны участники?*
>
> *У кого не было шансов присоединиться к панели?*
>
> *Кто с большей вероятностью участвовал и почему?*
>
> *Какие известные различия корректируют веса и какие они не корректируют?*
>
> *Было ли принято недействительное или поддельное решение о участии без просмотра содержимого ответа ИИ?*
>
> *Влияла ли награда участников на их поведение и выбор ответов?*

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Каждое наблюдение GEO-1000 должно иметь отслеживаемую цепочку выбора участников от целевой популяции до действительной аналитической записи; решения о выборе, приглашении, принятии, назначении заданий, выполнении, проверке, исключении и взвешивании должны быть независимы от положительного или отрицательного содержания ответа ИИ. [K04–K06] Панель участников только:

- страны,
- не подразумевает никакой пригодности в отношении возраста,
- язык

Не может считаться подходящей для представления популяции, даже если соблюдаются квоты. Панель также должна содержать следующую информацию в соответствующей мере:

- Источник участников
- Механизм выбора
- Метод участия в панели
- Критерии допустимости
- Уникальная проверка человека
- Проверка страны и языка
- Проверка доступа к продукту ИИ
- Информация об аккаунте и плане
- Устройство и интерфейс
- Частота использования ИИ
- Знакомство с брендом
- Отношения с проверяемой организацией
- Система стимулов
- Метод назначения заданий

- Отсутствие ответов
- Недействительность
- Повторение и панельная подготовка
- Взвешивание
- Отсутствие охвата
- Конфиденциальность и согласие

Измерение некоторой информации может повлиять на участника. Поэтому порядок сбора данных также является частью протокола.

3. ЦЕПОЧКА ВЫБОРА УЧАСТНИКОВ

Цепочка от целевой популяции до окончательного наблюдения состоит из следующих этапов:

U→F→S→I→E→A→C→V

Здесь:

- U: целевая пользовательская популяция
- F: лица, включенные в выборочную рамку
- S: выбранные люди
- I: приглашенные лица
- E: лица, чья правомочность была подтверждена и которые были включены в исследование
- A: люди, назначенные на задание
- C: люди, выполняющие задание
- V: люди, создающие корректные аналитические наблюдения

На каждом этапе перехода люди могут теряться или может появляться новый источник смещения. Например:

- Люди в целевой популяции, которые не находятся в панели
- Люди в панели, которые не были приглашены
- Приглашенные люди, которые не согласились
- Те, кто согласился, но не завершил задание
- Те, кто завершил, но создал некорректные записи

— это не одни и те же люди. Поэтому можно сообщать только о последних 1,000 корректных наблюдениях. Воронка отбора должна быть полностью видимой.

4. ИСТОЧНИКИ УЧАСТНИКОВ

GEO-1000 не обязательно должен быть привязан к единственному источнику участников. Каждый источник имеет свои собственные предвзятости и охват.

4.1. Панель вероятностной выборки населения

У участников есть известная и ненулевая вероятность быть отобранными из определенной популяции. Источники:

- домохозяйство,
- адрес,

- население или действительная запись,
- вероятностный телефон или многоступенчатый отбор

может быть. Преимущества:

- можно установить вес дизайна,
- выводы о популяции становятся более достоверными,

процесс отбора может быть более прозрачным. Ограничения:

- может быть дорогостоящим в глобальном масштабе,
- в некоторых странах может не быть подходящей рамки,
- также может понадобиться найти подгруппу населения с доступом к продуктам ИИ

Снова будет отсутствие ответа. Вероятностный выбор не является автоматически идеальным представлением.

4.2. Профессиональная панель добровольцев для исследований

Участники ранее добровольно зарегистрировались в исследовательской панели. Преимущества:

- это быстро,
- можно получить доступ к нескольким странам,
- можно предоставить профиль и знание языков,

можно проводить повторные измерения. Ограничения:

- участники склонны к исследованиям,
- уровень использования технологий может быть высоким,
- могут присутствовать активные пользователи панели,
- вероятности попадания в популяцию могут быть неизвестны,

один и тот же человек может быть зарегистрирован более чем на одной панели. Эта панель:

не может быть представлена как вероятностная выборка мирового населения

Это не должно рассматриваться как вероятностная выборка мировой популяции.

4.3. Открытый онлайн-призыв

Участие осуществляется через социальные сети, веб-сайт или открытую регистрационную форму. Преимущества:

- широкая видимость,
- низкая стоимость,
- доступ к небольшим странам и языковым сообществам

может быть обеспечено. Ограничения:

- сильная самоотборка,
- концентрация фанатов или критиков бренда,
- фальшивое и повторное участие,
- чрезмерное представление энтузиастов технологий

может возникнуть. Открытый призыв может быть гораздо более подходящим, чем основная Панель Населения для:

- обнаружение,
- Наблюдатель страны,

- Языковой справедливости,
- задачи репликации

Эти источники могут быть лучше приспособлены для этих целей.

4.4. Панель, связанная с университетом или учреждением

Участники:

- выбираются через университеты,
- государственным учреждением,
- исследовательские центры,
- профессиональные организации

Преимущества:

- местный доступ,
- языковая экспертиза,
- этический контроль,
- достижение сообществ с ограниченными ресурсами

может быть обеспечено. Ограничения:

- интенсивность студентов или профессионалов,
- эффект учреждения,
- возрастной и образовательный перекос,
- давление участия со стороны корпораций

может создавать.

4.5. Список клиентов или пользователей

Выбираются фактические клиенты или пользователи проверяемого учреждения. Этот источник:

- опыт клиента,
- пригодность для обслуживания,
- узнаваемость бренда

Ценен для измерений. Это не панель общей популяции. Клиенты:

- может лучше познакомиться с брендом,
- может получить более положительный или более отрицательный опыт,

может быть более склонен задать соответствующий запрос. Панель клиентов должна быть определена отдельно.

4.6. Участники, предложенные проверяемой организацией

Компания:

- сотрудник,
- клиент,
- партнер,
- подписчик

может предлагать. Эти пользователи не могут быть включены в основную независимую популяционную панель без объяснения. Отдельно:

ПАНЕЛЬ, НОМИНИРОВАННАЯ СУБЪЕКТОМ

должна быть помечена. Если организация выбирает участников, результат не является независимым измерением популяции.

4.7. Гибридная панель из нескольких источников

Это комбинация нескольких источников. Пример:

- национальная панель на основе вероятности,
- профессиональная онлайн-панель,
- местный партнер в небольшой стране,
- Сообщество языковой справедливости

может использоваться совместно. Гибридная панель может увеличить охват. Однако среди ресурсов:

- повторение индивидуумов,
- вероятности выбора,
- стимулы,
- опыт выполнения заданий,
- профиль пользователя

также должны быть смоделированы.

5. СТАТУСЫ ЦЕЛОСТНОСТИ ВЫБОРА УЧАСТНИКОВ

Качество выбора участников в панели может быть классифицировано следующими статусами.

PSI-0— ВЫБОР НЕИЗВЕСТЕН

Недостаточно известно, где и как были выбраны участники. Обобщение на популяцию невозможно.

PSI-1— УДОБНАЯ ИЛИ ОТКРЫТАЯ ВОЛОНТЕРСКАЯ ПОДПИСКА

Это удобная или открытая выборка добровольцев. Может использоваться для исследований и выявления событий.

PSI-2— УЧАСТИЕ С КОНТРОЛЕМ КВОТ

Существуют квоты по стране, языку или демографии. Вероятность попадания в популяцию неизвестна.

PSI-3— КАЛИБРОВАННАЯ ПАНЕЛЬ С МНОЖЕСТВОМ ИСТОЧНИКОВ

Существуют несколько источников, предопределенные квоты и калибровка популяции. Незмеренный систематический сдвиг при отборе может сохраняться.

PSI-4— ВЕРОЯТНОСТНЫЙ ИЛИ СИЛЬНЫЙ ГИБРИДНЫЙ ДИЗАЙН

Существует известная вероятность выбора или сильный гибридный дизайн с вероятностной и невероятностной выборкой.

PSI-5— РЕПЛИКИРОВАННЫЙ И ПРОВЕРЕННЫЙ С ВНЕШНИМ АУДИТОМ ОТБОР

Дизайн отбора:

- с разными поставщиками,
- в разные волны,
- под независимым аудитом

повторяющееся и предвзятое поведение было изучено. Эти статусы не гарантируют точность оценки. Они указывают на уровень уверенности, возникающий при отборе участников.

6. СЛУЧАЙНЫЙ ОТБОР ВНУТРИ ПАНЕЛИ И СЛУЧАЙНЫЙ ОТБОР ИЗ НАСЕЛЕНИЯ

В панеле добровольцев может быть 100,000 зарегистрированных пользователей. Из этой панели можно случайным образом выбрать 1,000 человек. Отбор:

> *осуществляется случайным образом внутри панели.*

Однако сама панель не формируется случайным образом из мировой популяции. Следовательно, нельзя сказать: «1,000 человек случайным образом из мировой популяции». Правильное выражение:

> *«1,000 участников, случайным образом выбранных внутри предварительно зарегистрированной панели добровольцев и калиброванных по целевой популяции.»*

Случайность должна ограничиваться этапом, на котором она применяется.

7. ЗАПИСЬ ОТНОСИТЕЛЬНО ПРАВОМОЧИЯ УЧАСТНИКА

Каждый участник должен соответствовать следующим условиям в той степени, в какой это уместно, прежде чем быть включенным в основной панель:

- Возрастная возможность участия
- Согласие добровольца
- Соответствие стране и местности
- Владение языком
- Доступ к продукту ИИ
- Правильный план или пользовательский интерфейс
- Необходимое устройство и технический доступ
- Уникальный человеческий статус
- Независимость выполнения задач
- Декларация о заинтересованности или взаимоотношениях
- Безопасность исследования
- Согласие на сбор данных и доказательств

Решение о соответствии не может быть изменено задним числом после того, как сгенерирован ответ ИИ.

8. ВЕРИФИКАЦИЯ СТРАНЫ

Вопрос «К какой стране относится участник?» не имеет однозначного ответа. Следующие поля должны быть разделены:

- Обычная страна проживания
- Физическая страна на момент измерения
- Страна или рыночный регион аккаунта ИИ
- Страна оплаты или плана
- Гражданство
- Целевая страна запроса
- Локаль

Гражданство может не быть основной переменной для большинства географических измерений. Что важно в большинстве случаев:

> *Это фактическое географическое и продуктовое состояние, в котором пользователь взаимодействует с продуктом.*

8.1. Путешествующий пользователь

Участник может находиться в стране, отличной от его обычной страны проживания. Этот человек:

- население по месту проживания,
- место измерения,
- путешествующий пользователь

должен иметь отдельный статус как.

8.2. VPN или смена местоположения

Использование VPN:

- конфиденциальность
- сетевая инфраструктура учреждения,
- обычные привычки пользователей

могут быть причиной. Это не следует рассматривать как автоматическое мошенничество. Тем не менее, это затрудняет проверку уровня страны. В основной панели контролируемой страны может быть правило отключать или регистрировать использование VPN. Изменения местоположения, используемые для обхода ограничений, не подходят для основной панели Естественный охват.

8.3. Уровни проверки страны

GV-0— Непроверенный

Доступна только декларация от пользователя.

GV-1— Подтверждено панелью

У провайдера панели уже есть регистрация страны.

GV-2— Последовательность сеансов

Декларация, локаль и сигналы сеанса последовательны.

GV-3— Многофакторное подтверждение

Защита конфиденциальности совместима с несколькими сигналами.

GV-4— Независимо проверено

Метод проверки страны прошёл внешнюю проверку качества. Сбор точного адреса или чувствительных координат не является обязательным требованием. Проверка страны должна проводиться с минимизацией данных.

9. ВЛАДЕНИЕ ЯЗЫКОМ И ЕГО ЕСТЕСТВЕННОЕ ИСПОЛЬЗОВАНИЕ

Знание языка участником может основываться исключительно на утверждении: «Да, я знаю». Это может быть низкорисковая первоначальная регистрация. В панели с высоким уровнем безопасности языка дополнительно:

- заявление о естественном использовании,
- короткая проверка понимания,
- история панели,
- оценка местным экспертом,
- нейтральное испытание перед заданием

может быть использовано. Языковой тест не должен превращаться в тяжёлый академический экзамен, отбирающий только высокообразованных людей. Цель:

> *Чтобы участник понимал промпт и ответ как настоящий пользователь.*

10. ПРОВЕРКА ДОСТУПА К ПРОДУКТУ ИИ

Участник:

- должен иметь возможность получить доступ к правильному продукту ИИ,
- правильный пользовательский интерфейс
- к правильному плану или состоянию учётной записи

должен иметь возможность доступа. Следующие ситуации должны фиксироваться отдельно:

- Бесплатный план
- Платный индивидуальный план
- Корпоративный план
- План обучения
- Веб
- Мобильный
- Настольный
- Встроенный продукт
- Видимая модель
- Веб/функция поиска

Участник не может использовать чужую учетную запись. Нельзя запрашивать совместное использование пароля. Доступ к продукту должен быть подтвержден без публичного раскрытия идентификатора учетной записи.

11. ОТНОШЕНИЯ МЕЖДУ УЧАСТНИКОМ И АУДИРУЕМОЙ ОРГАНИЗАЦИЕЙ

Отношение участника к контролируемой организации может влиять на поведение при ответах или знакомство с брендом. Статусы отношений:

ОБЩАЯ ПУБЛИКА

ТЕКУЩИЙ КЛИЕНТ

БЫВШИЙ КЛИЕНТ

ПЕРСПЕКТИВНЫЙ КЛИЕНТ

СОТРУДНИК

БЫВШИЙ СОТРУДНИК

КОНТРАКТОР

ПАРТНЕР

ИНВЕСТОР

КОНКУРЕНТ / АФФИЛИАТ

АФФИЛИАТ ПОСТАВЩИКА ИИ

ИССЛЕДОВАТЕЛЬ

НЕИЗВЕСТНО

Прямые связи с сотрудниками и лицами, принимающими решения, в основной панели общей популяции:

- могут быть исключены,
- могут быть помещены в отдельный слой,

может быть защищен с помощью низкого истинного весового коэффициента популяции. Автоматическое исключение всех связанных лиц также может быть неверным. Если они являются частью фактической клиентской или сотруднической популяции, их можно измерить отдельно и явно.

12. ЗНАКОМСТВО С БРЕНДОМ

Аудируемый участник:

- никогда о нем не слышал,
- знает название,
- знает продукт,
- является клиентом,

- может иметь сильное положительное или отрицательное мнение.

Эта информация:

- может влиять на поведение участника,
- Панель пользователей в естественных условиях,
- требуется немедленный ответ

желаем. Если знакомство измеряется перед запросом, это может подготовить пользователя к бренду. Следовательно, порядок предпочтений:

- Ранее существующая информация панели
- Отдельный вопрос после задания
- При необходимости нейтральное предварительное измерение, которое не направляет бренд

может быть проведено. Знакомство должно быть видно в основной выборке; его не следует выбирать в качестве весового переменного после того, как будут получены результаты.

13. ЗНАКОМСТВО С ИСПОЛЬЗОВАНИЕМ ИИ

Участники:

- первый раз пользователь
- редкий пользователь,
- обычный пользователь,
- интенсивный или профессиональный пользователь

люди могут быть. Частота использования ИИ:

- доступ к продукту,
- тип плана,
- знания интерфейса
- персонализация,
- вероятность правильного выполнения задачи

может быть затронуто. Тем не менее, выборку только опытных пользователей ИИ может снизить ошибки протокола. Но это может исказить оценку естественной популяции. Правильное решение:

- измерять уровень использования,
- сравните это с целевой популяцией,
- стратифицировать или взвешивать при необходимости,

и сохранять достоверный реальный опыт неопытного пользователя.

14. УНИКАЛЬНОЕ ПРАВИЛО ДЛЯ ЧЕЛОВЕКА

Единицей популяции GEO-1000 является человек. Один и тот же человек:

- в нескольких панелях,
- с несколькими учетными записями электронной почты,
- с несколькими учетными записями ИИ,
- с разных устройств

может участвовать. Проверка уникальности должна выполняться с сохранением конфиденциальности. Могут использоваться следующие сигналы:

- Анонимная идентификация участника
- Криптографическая или хеш-основанная повторная проверка
- Соответствие устройству и сессии
- Запись о платеже или вознаграждении
- Время и шаблон задания
- Один и тот же скриншот или файл ответа
- Безопасная повторная проверка через нескольких поставщиков панелей

Ни один отдельный сигнал не должен быть достаточным для автоматического исключения. В частности:

- Семьи, использующие одно общее устройство,
- сетевая инфраструктура учреждения,
- Инструменты доступности,
- Пользователи общих устройств с низким доходом

могут быть ошибочно признаны поддельными. Подозрительные записи должны проходить проверку человеком и процесс апелляции.

15. НАГРАДА УЧАСТНИКА

Оплата или вознаграждение, предоставляемое участнику, может влиять на их поведение относительно данных.

15.1. Оплата независимо от результата

Предпочтительный метод:

> *Фиксированная и заранее обьявленная оплата за выполнение задания согласно протоколу.*

Оплата:

- не может зависеть от того, является ли ответ положительным,
- не может быть вынуждено к категориям

  истинно,
- длинным,
- благоприятным для бренда,
- или сопровождаться цитатами.

Отбор должен оставаться независимым от всех таких исходов.

15.2. Бонус за качество

Если используется бонус за качество, только:

- полные доказательства,
- правильный запрос,
- правильный продукт,
- своевременная поставка,
- очистка персональных данных

может быть связана с условиями протокола. Она не может быть привязана к содержанию ответа ИИ.

15.3. Стимул за скорость

Бонус исключительно за скорость:

- торопливость,
- неполное захватывание,
- неправильный продукт,
- сокращение ответа

может повысить риск. В пределах установленного времени следует предоставлять разумный срок.

15.4. Оплата по странам

Оплата может варьироваться в зависимости от местных сборов и условий жизни. Эта ситуация:

- может повлиять на вероятность участия,
- состав панели

и должна быть зафиксирована вместе с подходом к оплате и различиями между странами.

15.5. Принуждение и чрезмерное стимулирование

Вознаграждение не должно создавать давления, от которого человек не может разумно отказаться. В частности:

- низкий доход,
- студент,
- зависимые сотрудники,
- уязвимые группы

должны иметь защищённое право на добровольность.

16. ТЕКСТ ПРИГЛАШЕНИЯ И ПРЕДВАРИТЕЛЬНАЯ ОРИЕНТАЦИЯ

Приглашая участника, цель исследования может подготовить его к определённому результату. Неправильное приглашение: «Мы тестируем, как AI-системы неправильно интерпретируют Apple». Это заявление может заставить участника искать ошибки. Ещё одно неправильное приглашение: «Помогите нам доказать успех нашего бренда с помощью AI». Это заявление создаёт ожидание положительного результата. Предпочтительное нейтральное приглашение: «Изучение стандартизированного пользовательского опыта и наблюдение за представлением в конкретных ИИ-продуктах». Аудируемый объект может стать видимым позднее из-за характера задачи. Тем не менее, желаемый результат не должен раскрываться.

17. НАЗНАЧЕНИЕ ЗАДАЧИ

Участники:

- Продукт ИИ,
- запрос,

- языка,
- волне,
- условие сессии,
- порядок выполнения задач

должны быть назначены заранее с использованием заранее определённого метода. Участник:

- любимый ИИ-продукт,
- план, который он/она знает лучше всего,
- язык с высокой вероятностью получения положительного отклика

если разрешено выбирать самостоятельно, может возникнуть смещение при назначении.

17.1. Случайное назначение в пределах слоя

После проверки страны, языка и доступа к продукту задачи могут быть назначены случайным образом внутри слоя.

17.2. Блочное назначение

Блоки могут использоваться для балансировки бесплатных/платных планов, мобильных/веб, или частоты использования ИИ.

17.3. Порядок продуктов

Если человек тестирует более одного продукта, порядок должен быть случайным. Ответ на первый продукт может повлиять на оценку следующего ответа или поведение при выполнении задачи.

17.4. Порядок промптов

Если участник получает более одной промпта:

- обучение,
- приобретение новой информации о бренде,
- перенос предыдущего ответа на следующий вопрос

возникает риск. Порядок промптов должен быть зафиксирован и при необходимости сбалансирован.

18. КОНДИЦИОНИРОВАНИЕ ПАНЕЛИ

Пользователи, которые часто участвуют в исследовательских панелях, могут узнать ожидания задания. Человек, который участвует в одном и том же тесте неоднократно:

- может запомнить правильный ответ,
- может более тщательно выбирать снимок экрана,
- может заранее просмотреть сайт проверяемого объекта,

можно узнать, какие ошибки ищутся. Эта ситуация называется:

Кондиционирование панели

Кондиционирование панели может изменять реальный естественный пользовательский опыт.

18.1. Источники кондиционирования

Предыдущие волны GEO Повторное измерение того же объекта Банк тех же промптов Тренировка по отзывам коллег Обмен заданиями в сообществах участников Предыдущие выплаты и обратная связь по качеству

18.2. Контроль кондиционирования

Статус нового и повторного участника
Дата первого участия
Предыдущее взаимодействие с тем же объектом
Предыдущее взаимодействие с той же семейством промптов
Ротационная панель
Случайность порядка заданий
Указанный период отдыха
Подвыборка новых пользователей
Сравнение результатов обученных и новых пользователей
Повторное использование одного и того же участника не запрещено. Это ценно для продольного измерения. Не может быть представлено как естественная поперечная панель.

19. КОНТАМИНАЦИЯ ЗАДАНИЙ И ОБЩЕНИЕ УЧАСТНИКОВ

Участники:

- промпты,
- ответы,
- скриншоты,
- ожидаемые результаты

могут быть переданы друг другу. Эта ситуация может нарушить независимые наблюдения. Контрольные меры:

- Ограничение продолжительности волны
- Случайный порядок заданий
- Обязанность не делиться результатами с участниками
- Поиск повторяющихся ответов и изображений
- Анализ необычных текстовых и временных совпадений
- Мониторинг предупреждений от поставщиков сообществ или панелей

Общение с участниками не является автоматически мошенничеством. Следует изучить его материальное влияние.

20. ВИДЫ НЕОТВЕТОВ

Неотвечание не является единичным событием.

20.1. Неответ на приглашения

Участник вовсе не отвечает на приглашение к исследованию.

20.2. Отказ от участия

Человек видит приглашение и решает не участвовать.

20.3. Не начало задания после квалификации

Человек признан подходящим, но не начинает задание.

20.4. Прерывание задания

Участник прекращает задание на полпути.

20.5. Не загрузка доказательств

Возможен ответ ИИ, но требуемая запись не загружена.

20.6. Неверное выполнение

Задача выполнена, но:

- неправильный запрос,
- неправильный продукт,
- выбранный ответ,
- отсутствует скриншот

являются причинами недействительности. Соотношения этих этапов должны сообщаться отдельно.

21. ВОРОНКА ОТВЕТОВ

Для слоя h:

- nhsel: выбранный человек
- nhinv: приглашенный человек
- nhcon: человек, согласившийся участвовать
- nhsta: человек, начавший задачу
- nhcom: человек, завершивший
- nhval: действительное наблюдение

Давайте посмотрим. Основные коэффициенты:

```
RR_h^invite= n_h^con/n_h^inv; RR_h^complete= n_h^com/n_h^sta; RR_h^valid= n_h^val/n_h^com
```

Это можно показать как. Каждый из этих показателей указывает на разную проблему. Указание только окончательного правильного значения скрывает смещение поля.

22. СМЕЩЕНИЕ ИЗ-ЗА ОТСУТСТВИЯ ОТВЕТОВ

Люди, которые не отвечают, могут отличаться от тех, кто отвечает. Пример: пользователь с медленным интернетом не может выполнить задачу. Пользователь с бесплатным тарифом не имеет доступа к продукту. Неподготовленный пользователь находит инструкции сложными. Активный пользователь ИИ выполняет задачу быстрее. В некоторых странах уровень оплаты остается низким. Пользователь с негативным опытом взаимодействия с брендом может быть более склонным к участию. Даже если уровень неответов высок, смещение может быть низким. Даже если уровень неответов низок, небольшая группа отсутствующих может иметь значительное финансовое значение. Поэтому важно учитывать не только сам уровень:

Какие характеристики отличаются между респондентами и нереспондентами должно быть исследовано.

23. ВЕСЫ НЕОТВЕТА

Если будет использоваться настройка без ответа, классы следует определить до того, как будут получены результаты. Потенциальные переменные:

- Страна
- Язык
- Возраст
- Частота использования ИИ
- План
- Интерфейс
- Источник панели
- Доступность
- Знакомство с брендом

Коэффициент неответа:

```
a_r^{NR}= (Σ_{j∈S_r} d_j)/ (Σ_{i∈R_r} d_i)
```

может быть рассчитан как. Однако эта корректировка опирается только на предположение внутри одного класса: что респонденты достаточно репрезентативны для неотвечающих. Это не решает проблему неучтенных различий.

24. МЕХАНИЗМЫ ОТСУТСТВИЯ ДАННЫХ

Отсутствие данных концептуально можно рассматривать в трех классах.

24.1. СЛУЧАЙНОЕ ОТСУТСТВИЕ ДАННЫХ

Отсутствие данных не связано существенно с результатом и характеристиками пользователей. На практике это может быть редкостью.

24.2. ОТСУТСТВИЕ ДАННЫХ, ОБЪЯСНЯЕМОЕ НАБЛЮДАЕМЫМИ ОСОБЕННОСТЯМИ

Дефицит:

- страны,
- не подразумевает никакой пригодности в отношении возраста,
- план,
- язык

связан с хорошо известными областями, такими как. Взвешивание может частично помочь.

24.3. ДЕФИЦИТ, СВЯЗАННЫЙ С НЕСОБСТВЕННЫМИ ОСОБЕННОСТЯМИ

Например:

- Недоверие к ИИ-продукту,
- низкие цифровые навыки,
- противопоставление бренду,

- опасения по поводу конфиденциальности

если не измерять, взвешивание не может полностью исправить недостаток. В этом случае необходим анализ чувствительности.

25. ПОТЕРЯ ПАНЕЛИ И РАЗДЕЛЕНИЕ МЕЖДУ ВОЛНАМИ

В продольных или вращающихся панелях, где одни и те же пользователи измеряются снова, некоторые участники не принимают участие в следующей волне. Это потеря панели. Оставшиеся в панели:

- более дисциплинированы,
- более мотивированы наградами,
- более опытны,
- более привыкли к исследованиям

могут быть. Изменение оценки со временем может частично происходить из-за состава панели.

25.1. Запись об оттоке панели

В каждой волне:

- повторно приглашён,
- участвовал,
- вышел из участия,
- ново добавлен

должно отображаться количество пользователей.

25.2. Вес сокращения панели

Оставшиеся участники могут быть взвешены в соответствии с вероятностью выхода. Этот метод не полностью устраняет смещённость из-за необнаруженного выхода из панели.

25.3. Обновить панель

Новых пользователей можно добавить для замены потерянных. Результаты для новых и возвращающихся пользователей должны быть представлены отдельно.

26. МОШЕННИЧЕСТВО И АВТОМАТИЗАЦИЯ УЧАСТНИКОВ

Существует риск мошенничества, когда GEO-1000 является стимулируемой и глобальной панелью. Возможные действия:

- Человек, участвующий с несколькими аккаунтами
- Использование автоматических ботов
- Загрузка скриншота другого человека
- Редактирование изображений
- Загрузка одного и того же ответа от имени разных пользователей

- Указание неправильной страны или языка
- Совместное использование учетной записи
- Выполнение задания другим человеком
- Повторное присоединение среди поставщиков панелей
- Создание массовых ферм заданий

Этот риск не позволяет считать всех участников с низким доходом или быстрых участников подозрительными. Обнаружение мошенничества:

- обсуждения,
- множественные сигналы,
- соразмерная проверка человеком

должны быть основаны на этом.

27. СТАТУСЫ МОШЕННИЧЕСТВА И ИНТЕГРИТЕТА

FI-0— НЕ ОЦЕНЕНО

Запись не была оценена с точки зрения целостности.

FI-1— МАТЕРИАЛЬНЫХ АНОМАЛИЙ НЕТ

Явных признаков мошенничества не выявлено. Это не является абсолютной гарантией точности.

FI-2— ТРЕБУЕТСЯ ПРОВЕРКА

Обнаружен один или несколько подозрительных сигналов.

FI-3— ПОДОЗРЕНИЕ НА ДУБЛИКАТ ИЛИ СТРАХОВКУ

Есть подозрение на повторяющегося человека, визуальный, ответный или координационный сигнал.

FI-4— ПОДТВЕРЖДЕНА МАНИПУЛЯЦИЯ ПРОТОКОЛОМ

Участник намеренно изменил протокол.

FI-5— ПОДТВЕРЖДЕНО ИСПЫТОВОЕ НАБЛЮДЕНИЕ

Наблюдение не является подлинным опытом пользователя ИИ. FI-2 не является автоматическим исключением. Требуется проверка.

28. МОШЕННИЧЕСКИЕ СИГНАЛЫ

Одинаковый хэш скриншота Одинаковый редкий файловый след Невозможное время выполнения Многократное повторение идентичностей с одного и того же устройства или сеанса Конфликт материалов по стране, региону и доступу к продукту Несоответствие между текстом ответа и изображением Следы редактирования изображения Один и тот же человек в основном и резервном списках Один и тот же анонимный человек у нескольких поставщиков панелей Серийное и механическое выполнение задач Одинаковая необычная ошибка Одиночный сигнал может не быть убедительным доказательством мошенничества.

29. СПРАВЕДЛИВОСТЬ В ОБНАРУЖЕНИИ МОШЕННИЧЕСТВА

Следующие пользователи могут быть ошибочно отмечены как мошенники:

- Семьи, использующие общее устройство
- Сотрудники, использующие корпоративную сеть
- Участники университетских лабораторий
- Люди, заботящиеся о конфиденциальности, использующие VPN
- Пользователи с ограниченными возможностями, использующие экранные читалки или вспомогательные технологии
- Те, кто загружает похожие сжатые изображения из-за низкой пропускной способности

Таким образом:

- автоматическая маркировка,
- человеческая проверка,
- описание участника при необходимости,
- апелляция

должны работать совместно.

30. ВАЛИДНОСТЬ УЧАСТНИКА ДОЛЖНА БЫТЬ ОТДЕЛЬНА ОТ ТОЧНОСТИ ОТВЕТА ИИ

Запись участника может соответствовать протоколу. Ответ ИИ может быть неверным. Это действительное наблюдение. Ответ ИИ другого участника может быть верным. Однако пользователь:

- обновил ответ три раза,
- загрузил только лучший ответ,
- возможно, изменил запрос

Это недействительное наблюдение. Следовательно:

> *Правильный ответ ИИ не означает действительные данные. / Неправильный ответ ИИ не означает недействительные данные.*

Оценка валидности должна проводиться до основной оценки и быть максимально слепой.

31. КОНФЛИКТ ИНТЕРЕСОВ И ИСТОЧНИК УЧАСТНИКА

Участник или поставщик панели могут извлечь выгоду из следующих исходов:

- Аудитируемая компания получает высокий балл
- ИИ-продукт выглядит лучше
- Конкурент получает низкий балл
- Продажа большего количества аудиторских заданий
- Награда или трудовые отношения

Таким образом:

- отношения с участниками,
- соглашение с поставщиком панели,
- оплата труда, не зависящая от результатов,
- Вмешательство клиента

должно быть объяснено. Аудитируемое учреждение:

- кого из пользователей приглашать,
- кто будет считаться недействительным,
- Кого заменят резервным участником

не может решать.

32. ЭФФЕКТ ПОСТАВЩИКА ПАНЕЛИ

Пользователи двух поставщиков панелей в одной и той же стране могут различаться. Поставщик:

- Энтузиасты технологий,
- студенты,
- участники интенсивных исследований

могут иметь более высокий уровень вовлеченности. Поставщик панели должен учитываться как кластерная или исходная переменная в анализе. Если возможно:

- результаты по поставщикам,
- эффект источника,
- сравнение поставщиков в рамках одной страны

должны быть исследованы. Глобальная панель с одним поставщиком не может представляться как мировое население.

33. ПОВТОРИТЬ ЛИЦО В МНОГОИСТОЧНИКОВОЙ ПАНЕЛИ

Один и тот же человек:

- две профессиональные панели,
- в открытом наборе,
- в списке университета

можно найти. Многоисточниковая панель может увеличить количество записей без увеличения уникальности. Проверка на дублирование между источниками:

- сохранение конфиденциальности,
- не требующая раскрытия центральной идентификации,
- позволяющая возражать против несоответствий

должна быть установлена соответствующим образом.

34. КАРТА СМЕЩЕНИЯ ВЫБОРКИ

GEO-1000 следует отдельно отслеживать следующие виды смещения.

SB-1— Смещение охвата

Некоторые сегменты целевой популяции отсутствуют в рамке выборки.

SB-2— Смещение отбора

Вероятность выбора отдельных лиц из рамки для участия в панели не соответствует запланированному дизайну.

SB-3— Смещение самовыбора

Добровольцы для исследования систематически отличаются от популяции.

SB-4— Смещение из-за приглашения и инструкций

Текст приглашения или источник привлекает людей с определенной точкой зрения.

SB-5— Смещение из-за отсутствия ответа

Люди, которые не участвуют или не завершают участие, отличаются по результатам.

SB-6— Смещение оттока участников панели

Пользователи, остающиеся между волнами, отличаются от тех, кто выпадает.

SB-7— Смещение из-за вознаграждения

Оплата или награда влияет на скорость, поведение при выборе или ответы участника.

SB-8— Панельное кондиционирование

Участник усваивает ожидаемое поведение из предыдущих исследований.

SB-9— Смещение при назначении заданий

Участники сами выбирают продукт, язык или промпты или назначаются неравномерно.

SB-10— Смещение классификации по стране и языку

Участнику назначена неправильная страна или языковая группа.

SB-11— Смещение продукта, плана и интерфейса

Только определенные пользователи плана или устройства определяют панель.

SB-12— Смещение по знанию бренда

Клиенты, фанаты, критики или сотрудники представлены в избытке.

SB-13— Смещение из-за мошенничества и повторного участия

Один и тот же человек или автоматическое участие получает слишком большое влияние на результат.

SB-14— Смещение из-за исключения по валидности

Отрицательные ответы ИИ чаще считаются недействительными.

SB-15— Смещение при взвешивании

После результата выбранные переменные или чрезмерное изменение веса модели направляются в желаемое направление.

SB-16— Смещение оставшейся выборки

Представлены только пользователи, которые могут успешно выполнить задачу, как в реальной популяции продукта.

35. ЦЕЛЬ ВЗВЕШИВАНИЯ

Взвешивание может служить следующим целям:

- Корректировка неравных вероятностей отбора
- Приведение избыточной выборки к фактической пропорции в популяции
- Снижение различий в уровне откликов
- Калибровка выборки по известным целевым суммам
- Объединение нескольких источников панелей

Взвешивание не может служить следующим целям:

- Создание человека, которого нет в выборочной совокупности
- Точная корректировка неизмеренных переменных
- Превращение ложного или ошибочного наблюдения в реальное
- Получение желаемого результата после его появления
- Преобразование одной малой ячейки в надежный показатель по стране
- Автоматическое превращение непараметрической панели в вероятностную выборку

36. ТЯЖЁЛАЯ АРХИТЕКТУРА

Окончательный вес можно представить следующим образом:

```
w_i= d_i× a_i^{NR}× g_i^{CAL}× p_i^{SRC}
```

Здесь:

- di: базовый дизайн или начальный вес
- aiNR: настройка ответа
- giCAL: калибровка популяции
- piSRC: источник панели или корректировка маловероятной выборки

В вероятностной панели:

```
d_i= 1/π_i
```

Это возможно. Истинное πi неизвестно в непредставительной панели. В этом случае используемый вес не является весом дизайна, а является весом на основе модели или калибровочным весом. Это различие следует сохранить в отчете.

37. ВЕСЫ НЕ МОГУТ ЗАВИСЕТЬ ОТ РЕЗУЛЬТАТА

Следующие переменные нельзя использовать для создания весов:

- Является ли ответ AI правильным или нет
- Упоминается ли бренд или нет
- Есть ли ссылка
- Дается ли рекомендация

- Произошла ли критическая ошибка
- Нравится ли результат проверяемой организации

Взвешивающие переменные:

- До того, как будет виден результат,
- Потому что это связано с целевой популяцией и механизмом отбора

Она должна быть выбрана. Если результат ответа определяет вес, балл формируется сам собой.

38. ПЕРЕМЕННЫЕ КАЛИБРОВКИ

Возможные области калибровки:

- Страна
- Регион
- Язык
- Возраст
- Пол, если необходимо и этично
- Образование
- Городское/сельское поселение
- Тип интернет-доступа
- Устройство
- Частота использования ИИ
- Бесплатный/платный план
- Интерфейс
- Статус доступности

Надежные целевые показатели могут быть недоступны для каждой переменной. Слишком много переменных калибровки:

- может создавать чрезмерное разнообразие весов,
- низкую эффективную выборку
- переобучение модели

Её следует заморозить до того, как будут получены результаты калибровочного набора.

39. ВЗВЕШИВАНИЕ И СБАЛАНСИРОВАННОСТЬ

Образец до и после взвешивания следует сравнить с целевой популяцией. Для каждого поля:

- целевая доля
- доля сырого образца
- взвешенная доля
- оставшаяся разница

должны быть показаны. Взвешивание применяется только к количеству результатов, и таблица баланса не может быть сохранена.

40. ВЕСОВЫЕ ДИАГНОЗЫ

Основной отчет должен включать следующие поля, если они имеют отношение:

- Минимальный вес
- Максимальный вес
- Медианный вес
- Коэффициент вариации веса
- Соотношение весов
- Количество обрезанных весов
- Размер исходной выборки
- Эффективный размер выборки
- Эффект дизайна
- Невзвешенный результат
- Основной взвешенный результат
- Альтернативные результаты с разумными весами

Взвешенное числовое значение не является полной контрольной записью без идентификаторов веса.

41. ОТОБРАЖЕНИЕ СИНТЕТИЧЕСКОГО ВЕСА

ОТОБРАЖЕНИЕ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Участники, показатели и оценки ниже гипотетические. Целевая популяция и распределение по панелям:

Частота использования ИИ	Доля целевой популяции	Доля исходной панели	Исходное наблюдение
Редкое или низкое использование	50%	10%	100
Среднее использование	30%	30%	300
Интенсивное использование	20%	60%	600
Итого	100%	100%	1,000

Синтетические показатели прохождения:

Группа	Доля успешных результатов	Предыдущий ответ
Редкое или низкое использование	72%	72
Среднее использование	84%	252
Интенсивное использование	96%	576
Итого	90%	900

Исходный результат:

```
900/1000=90%
```

41.1. Калибровочные коэффициенты

Низкое использование:

```
0.50/0.10=5
```

Среднее использование:

```
0.30/0.30=1
```

Высокое использование:

```
0.20/0.60 = 1/3
```

Взвешенный ответ:

```
72(5) + 252(1) + 576(1/3) = 804
```

Сумма весов:

```
100(5) + 300(1) + 600(1/3) = 1,000
```

Взвешенный результат:

```
804/1000=80.4%
```

41.2. Эффективный размер выборки

```
n_eff= [100(5)+300(1)+600(1/3)]² / [100(5 ²)+300(1 ²)+600(1/3)²] ≈ 349
```

Карта результатов:

- Число действительных наблюдений: 1,000
- Исходный балл: 90 процентов
- Взвешенный балл: 80.4 процентов
- Эффективная выборка: приблизительно 349
- Основная причина: чрезмерное представительство активных пользователей ИИ

Правильная интерпретация:

> *В панели содержится 1,000 действительных наблюдений; однако в результате калибровки целевой популяции оценка снизилась до 80.4%, а из-за изменчивости весов эффективный размер выборки стал приблизительно 349.*

Неверный комментарий: «Глобальная точность была доказана на уровне 90% с участием 1,000 человек.»

42. ЧТО НЕ МОЖЕТ ИСПРАВИТЬ ВЗВЕШИВАНИЕ?

Следующие группы вообще не должны присутствовать в синтетической панели:

- Люди, использующие программы чтения с экрана
- Пользователи из сельских районов с низкой пропускной способностью
- Люди, использующие продукт AI впервые
- Малые языковые сообщества

Взвешивание только изменяет вес пользователей в панели. Не наблюдаемой группы:

- AI ваш ответ,
- выявить проблемы с доступом,
- распределение ошибок

не может быть получено. Поэтому:

> *Нулевая охватность означает отсутствие информации.*

Взвешенный результат: это может быть предсказание, более близкое к целевой популяции, а не полное наблюдение за целевой популяцией.

43. АНАЛИЗ ЧУВСТВИТЕЛЬНОСТИ

Необходимо изучить зависимость взвешенного результата от одной модели веса. Возможные результаты:

- Необработанный невзвешенный результат
- Результат, ориентированный на дизайн
- Результат, откалиброванный по демографии
- Калибровка с учётом использования ИИ
- Результат с усечённым весом
- Результат с учётом влияния поставщика панели

Если результаты изменяются существенно:

Предупреждение о чувствительности к весам

должно быть выдано. Нельзя представить один «правильный» результат с весом как абсолютную истину.

44. КОРРЕКТИРОВКА СКЛОННОСТИ В НЕВЕРОЯТНОСТНОЙ ПАНЕЛИ

Склонность участников невероятностной панели присоединяться к панели может быть смоделирована в соответствии с целевой популяцией. Вероятность того, что человек будет в панели:

```
p̂_i= Pr(i ∈ panel | X_i)
```

может быть оценено. Кандидат на обратный вес склонности:

```
w_i^P= 1/p̂_i
```

может быть использован. Этот метод:

- использует только измеренные переменные Xi,
- подвержен влиянию некорректно заданной модели,
- не решает проблему необнаруженного отбора,

требует достоверных эталонных данных. Поэтому результат:

должен быть обозначен как:

Модельно-калиброванная оценка непараметрической панели

45. ОЦЕНКИ МАЛЫХ ОБЛАСТЕЙ И НА ОСНОВЕ МОДЕЛИ

Прямое выборочное обследование может быть недостаточным в некоторых странах или языковых ячейках. Методы, основанные на моделировании, могут заимствовать информацию у соседних или схожих групп. Это означает, что результат:

- непосредственно наблюдаемый показатель по стране,
- сырой пользовательский рейтинг

не является таковым. Правильный статус:

ОЦЕНКА С ПОДДЕРЖКОЙ МОДЕЛИ

ОЦЕНКА МАЛЫХ ТЕРРИТОРИЙ

НЕ НАПРЯМУЮ НАБЛЮДАЕМОЕ

обязательно. Прогноз на основе модели не превращает страну без наблюдения Country Observer в наблюдаемую страну.

46. КОНФИДЕНЦИАЛЬНОСТЬ И ВОЗМОЖНОСТЬ ПРОВЕРКИ УЧАСТНИКА

GEO-1000 требует доказательства для каждого наблюдения. Это не означает, что личность участника будет разглашена публично. Снимок экрана может включать следующую информацию:

- Имя
- Электронная почта
- Изображение профиля
- Предыдущий чат
- План учетной записи
- Название учреждения
- Конфиденциальная личная информация

NOMOS Система захвата или аналогичная система:

- не должна захватывать ненужные персональные поля,
- должна выполнять автоматическое или контролируемое редактирование,
- исходные доказательства должны храниться в защищенном хранилище

только анонимные и проверяемые записи должны быть представлены общественности.

46.1. Манифест общественных доказательств

Общественная запись:

- ID наблюдателя,
- ячейка страна/язык,
- Продукт ИИ,
- время,
- хеш доказательства,
- статус проверки

может быть перенесен. Полный скриншот:

- отредактирован,
- доступ ограничен,
- открыт для независимого аудитора

может быть. Наличие тысячи доказательств не требует публичного раскрытия личности тысячи людей.

46.2. Риск чувствительных маленьких стран

Если в маленькой стране только один участник:

- устройство,
- время,
- языка,
- план

Когда данные собираются вместе, может стать легче повторно идентифицировать человека. Уровень детализации в публичных отчетах должен быть снижен. Целостность доказательств может поддерживаться в системе с контролем доступа.

47. СОГЛАСИЕ УЧАСТНИКА

Участник должен как минимум знать следующее:

- Какое задание он будет выполнять
- Какие данные и снимки экрана будут собраны
- Какая личная информация не будет собрана
- Как долго данные будут храниться
- Кто будет иметь доступ
- Как будут опубликованы результаты
- Как будет выдаваться вознаграждение
- Что участие является добровольным
- Право на отзыв или возражение
- Пароль от аккаунта запрашиваться не будет
- Не ожидается нарушение условий использования продуктов AI

Согласие не должно скрываться в длинном и непонятном тексте. [K06; K22]

48. БЕЗОПАСНОСТЬ УЧАСТНИКА

От участника:

- предоставлять личную медицинскую, юридическую или финансовую информацию ИИ,
- использовать реальные данные клиентов,
- делиться конфиденциальной информацией работодателя,
- ослаблять безопасность аккаунта,
- превышать лимиты использования,
- нарушать законы или условия использования продукта

не могут быть запрошены. Тестовые промпты должны максимально иметь публичный и неперсональный контекст.

49. УЧАСТНИК ПРЕДСТАВЛЯЕТ ЧЕЛОВЕЧЕСКОЕ РАЗНООБРАЗИЕ, А НЕ КАЧЕСТВО

Неопытный пользователь:

- может выполнять задачу медленнее,
- труднее,
- с большим количеством поддержки

Этот человек не является автоматически участником низкого качества. Они являются реальной частью целевой популяции. Если исследовательский дизайн настолько сложен, что его могут пройти только люди с высокими цифровыми навыками: проблема может быть не в участнике, а в инструменте измерения. Протокол:

- должен быть понятным,
- доступным,
- лингвистически естественным,
- техническая нагрузка ограничена

должен быть.

50. ОБЩЕСТВЕННАЯ КАРТА РЕЗУЛЬТАТОВ ВЫБОРА УЧАСТНИКА

Кандидатский общественный отчет должен включать следующие области:

Ресурсы панели

Доля панели по вероятности Доля профессиональной добровольной панели Доля открытого набора Доля университета/организации Доля клиентской или аффилированной панели

Воронка участия

Человек в кадре Выбран Приглашен Принял Начал Завершил Дает валидное наблюдение

Право на участие и честность

Уровни проверки страны Уровни проверки языка Частота повторного участия Подозрительная частота регистрации Проверенная частота мошенничества Частота недействительных наблюдений

Состав панели

Частота использования ИИ План и интерфейс Знание бренда Отношения с клиентами/сотрудниками Новые и возвращающиеся пользователи панели Покрытие доступности

Вес Диагнозы

Следующее может публиковаться без раскрытия идентичности участников: исходные и взвешенные показатели; минимальные и максимальные веса; коэффициент вариации весов; эффективный размер выборки; чувствительность к взвешиванию; и предупреждения о покрытии.

51. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH08-N01

Каждое действительное наблюдение GEO-1000 должно иметь прослеживаемую цепочку выбора участников от целевой популяции до аналитической записи.

CH08-N02

Случайный выбор внутри панели не может быть представлен как случайный выбор из целевой мировой популяции.

CH08-N03

Источник участников, метод отбора и статус панели должны быть объяснены в публичной записи метода.

CH08-N04

Не может быть определено как маловероятная или добровольная панельная выборка населения.

CH08-N05

Участники, предложенные проверяемой организацией, не могут быть включены в основную независимую панель населения без объяснения.

CH08-N06

Отношения с клиентами, сотрудниками, партнерами, конкурентами и поставщиками ИИ должны записываться отдельно.

CH08-N07

Критерии участия участников должны быть проверены до того, как будет увиден ответ ИИ.

CH08-N08

Страна, язык, продукт, план и интерфейс должны быть определены до назначения соответствующей задачи.

CH08-N09

Один и тот же человек не может дублироваться как единица населения из-за нескольких аккаунтов, устройств, языков, панелей или продуктов ИИ.

CH08-N10

Проверка уникальности личности должна проводиться в соответствии с принципами минимизации данных и конфиденциальности.

CH08-N11

Один технический сигнал сам по себе не может считаться достаточным для автоматического окончательного решения о том, что участник фиктивный.

CH08-N12

Вознаграждение участника не может быть связано с тем, является ли ответ ИИ положительным, правильным, длинным, содержащим ссылки или полезным для бренда.

CH08-N13

Бонус за качество может быть связан только с заранее определённой целостностью протокола.

CH08-N14

Текст приглашения и информация не могут направлять участника на получение положительного или отрицательного результата.

CH08-N15

Продукт ИИ должен быть назначен или рандомизирован до того, как будет виден результат последовательности промптов и задач.

CH08-N16

Выбор продукта и промпта на основе предпочтений участника не может использоваться необъяснённо в основной панели населения.

CH08-N17

Необходимо фиксировать условие панели, предыдущий опыт тестирования и повторное участие.

CH08-N18

Результаты новых и подготовленных участников должны анализироваться отдельно, когда это необходимо.

CH08-N19

Включение участника в панель, его сохранение, исключение или замена резервным участником не могут основываться на содержании ответа ИИ.

CH08-N20

Некорректные, отрицательные, отклоненные или критические ответы ИИ должны быть сохранены в соответствии с действующим протоколом.

CH08-N21

Действительность участника следует оценивать максимально слепым способом и до проверки фактической точности ответа.

CH08-N22

Уровни приглашений, принятия, начала, завершения и численно действительных наблюдений должны сообщаться отдельно.

CH08-N23

Настройка ответа должна основываться только на заранее определенных и релевантных переменных.

CH08-N24

Взвешивание не может представлять пользователей, которые отсутствуют в выборочной совокупности.

CH08-N25

Весовые переменные не могут зависеть от результата ответа ИИ, оценки или коммерческого интереса регулируемого учреждения.

CH08-N26

Весовые коэффициенты непараметрической панели не могут представляться как проектные веса.

CH08-N27

Калибровочные переменные и целевые итоги должны быть опубликованы до того, как будут видны результаты.

CH08-N28

Невзвешенные и основные взвешенные результаты должны храниться вместе.

CH08-N29

Диагностика взвешивания и эффективный размер выборки должны быть видны в основном отчёте.

CH08-N30

Если используется усечение веса, необходимо сохранить неусечённый результат и анализ чувствительности.

CH08-N31

Эффекты поставщика панели и источника участников должны храниться как отдельные переменные.

CH08-N32

В панели с несколькими источниками следует проводить проверки повторяющихся лиц между источниками.

CH08-N33

Обнаружение мошенничества должно включать множественные сигналы, пропорциональный человеческий контроль и процедуру апелляции.

CH08-N34

Использование VPN, общего устройства, сетей организаций или вспомогательных технологий не должно автоматически считаться мошенничеством.

CH08-N35

Модельная оценка малых областей не может быть представлена напрямую, как наблюдаемая оценка по стране или языку.

CH08-N36

Скриншоты участников и доказательства должны содержать минимизацию персональных данных, их редактирование и контроль доступа.

CH08-N37

Ни один участник не может быть запрошен предоставить пароли от аккаунта, историю личных чатов или ненужную конфиденциальную информацию.

CH08-N38

Участников нельзя просить обойти ограничения продукта, возраста, личности, страны или аккаунта.

CH08-N39

Дети или пользователи, нуждающиеся в особой защите, не могут быть зарегистрированы в панели без отдельных этических и безопасностных протоколов.

CH08-N40

Записи о выборе и взвешивании участников должны иметь ответственного человека или организацию.

52. ФОРМЫ ОШИБОК

CH08-F01— РАСЧЕТ СЛУЧАЙНЫХ УЧАСТНИКОВ В ПАНЕЛИ КАК СЛУЧАЙНЫХ ЛЮДЕЙ В МИРЕ

Случайный выбор внутри добровольной панели представлен так, как будто это случайный выбор из мировой популяции.

CH08-F02— РАСЧЕТ СООТВЕТСТВИЯ КВОТ КАК ПРЕДСТАВИТЕЛЬНОСТИ

Утверждается, что все поведенческие и технологические смещения устранены, поскольку соблюдены возрастные и страновые квоты.

CH08-F03— РАСЧЕТ КЛИЕНТОВ КАК ОБЩЕЙ ПОПУЛЯЦИИ

Список пользователей контролируемого учреждения используется так, как если бы это была мировая панель пользователей.

CH08-F04— СКРЫТИЕ СОТРУДНИКОВ И ПАРТНЕРОВ

Пользователи, связанные с учреждением, отображаются как независимые общие участники.

CH08-F05— ПРИСЧИТЫВАНИЕ ОДНОГО ПОСТАВЩИКА ПАНЕЛИ К ГЛОБАЛЬНОМУ

Эффект поставщика и поведение зарегистрированной панели не исследуются.

CH08-F06— ПРИСЧИТЫВАНИЕ ОТКРЫТОГО ПРИГЛАШЕНИЯ К ВЕРОЯТНОЙ ВЫБОРКЕ

Самоотбор сохраняется.

CH08-F07— ОПРЕДЕЛЕНИЕ СТРАНЫ ТОЛЬКО ПО IP

Различия, связанные с поездками, VPN, сетями учреждений и регионом учетной записи, игнорируются.

CH08-F08— ПРИСЧИТЫВАНИЕ ГРАЖДАНСТВА К СТРАНЕ ИСПОЛЬЗОВАНИЯ

Фактическое географическое положение опыта использования продукта неверно классифицировано.

CH08-F09— ПРИСЧИТЫВАНИЕ ЗАЯВЛЕННОГО ЯЗЫКА К КОНЕЧНОМУ, БЕЗ ТЕСТИРОВАНИЯ

Участник остаётся в языковой ячейке, даже если он не понимает промпт или ответ.

CH08-F10— ПРЕВРАЩЕНИЕ ЯЗЫКОВОГО ТЕСТА В ЭЛИТНЫЙ ОТБОР

На панели остаются только высокообразованные люди.

CH08-F11— ПОДСЧЁТ КОЛИЧЕСТВА АККАУНТОВ КАК ЛЮДЕЙ

Один человек увеличивает вес населения с помощью нескольких аккаунтов.

CH08-F12— ПОДСЧЁТ ОБЩЕГО АККАУНТА КАК ОДНОГО ЧЕЛОВЕКА

Аккаунт, используемый несколькими людьми, интерпретируется как один пользователь.

CH08-F13— ОПЛАТА В ЗАВИСИМОСТИ ОТ РЕЗУЛЬТАТА

Бонус выдается за положительный, правильный или ссылающийся ответ.

CH08-F14— НАРУШЕНИЕ ПРОТОКОЛА С ПРИЕМОМ БОНУСА ЗА СКОРОСТЬ

Участник делает скриншот или спешит через задание до завершения ответа.

CH08-F15— ОБЪЯСНЕНИЕ РЕЗУЛЬТАТА В ПРИГЛАШЕНИИ

Участнику предлагается найти ошибку или успех.

CH08-F16— ПОЗВОЛЕНИЕ УЧАСТНИКУ ВЫБИРАТЬ ПРОДУКТ

Пользователи, которые выбирают только продукт ИИ, к которому имеют доступ или который им нравится, определяют панель.

CH08-F17— СКРЫТИЕ ВЛИЯНИЯ ПОРЯДКА промптов

Предыдущие ответы влияют на последующие задачи.

CH08-F18— РАССМОТРЕНИЕ КОНДИЦИОНИРОВАНИЯ ПАНЕЛИ КАК ЕСТЕСТВЕННОГО ИСПОЛЬЗОВАНИЯ

Пользователь, который видит один и тот же тест несколько раз, представляется как новый пользователь.

CH08-F19— ИГНОРИРОВАНИЕ РАССЫЛКИ РЕЗУЛЬТАТОВ

Участники передавали задачи и ответы друг другу.

CH08-F20— УЧЁТ НЕОТВЕЧЕННЫХ ВОПРОСОВ КАК НЕДОСТАЮЩИХ ДАННЫХ

Систематическая разница между не участвующими не изучается.

CH08-F21— УЧЁТ ПОТЕРИ ПАНЕЛИ КАК УЛУЧШЕНИЕ РЕЗУЛЬТАТОВ

Оценки растут, потому что слабые или неопытные пользователи выпадают на последующих волнах.

CH08-F22— ОТМЕНА ОТРИЦАТЕЛЬНЫХ ОТВЕТОВ

Участник заменяется, потому что ИИ дал неправильный ответ.

CH08-F23— ПОДТВЕРЖДЕНИЕ ПРАВИЛЬНОГО ОТВЕТА

Даже если участник обновил ответ, запись сохраняется, потому что результат правильный.

CH08-F24— ИГНОРИРОВАНИЕ БОТОВ И ПОВТОРНОГО УЧАСТИЯ

Один и тот же человек или автоматизация создают большое количество наблюдений.

CH08-F25— МАССОВОЕ ИСКЛЮЧЕНИЕ С ПОМОЩЬЮ ОДНОГО СИГНАЛА

VPN, устройство или скорость делают так, что настоящие пользователи считаются фейковыми.

CH08-F26— ПРОВЕДЕНИЕ ПЕРЕВЕРКИ МОШЕННИЧЕСТВА НА ОСНОВЕ РЕЗУЛЬТАТОВ

Записи с низкими оценками проверяются чаще.

CH08-F27— СОЗДАНИЕ СИЛЬНО НЕВИЗУАЛИЗИРОВАННОЙ ГРУППЫ

Утверждается, что пользователи, отсутствующие в выборке, представлены.

CH08-F28— ВЫБОР КАЛИБРОВКИ ПОСЛЕ РЕЗУЛЬТАТА

Переменные веса, которые дают наиболее подходящую оценку, выбираются позже.

CH08-F29— ПОДСЧЕТ СЫРОЙ ДОЛИ ПАНЕЛИ КАК ДОЛИ НАСЕЛЕНИЯ

Активные пользователи ИИ определяют глобальный результат по сырым числам.

CH08-F30— ПОДСЧЕТ НЕОБЫЧНОГО ВЕСА КАК КОНСТРУКЦИОННОГО ВЕСА

Точное определение популяции выполняется, даже если истинная вероятность выбора неизвестна.

CH08-F31— СКРЫТИЕ НЕВЗВЕШЕННОГО РЕЗУЛЬТАТА

Эффект модели веса не виден.

CH08-F32— СКРЫТИЕ ЭФФЕКТИВНОЙ ВЫБОРКИ

Хотя сырые записи 1,000 содержат 300–400 единиц информации, заявлена полная точность.

CH08-F33— УДАЛЕНИЕ СЛАБЫХ ГРУПП ПО ВЕСОВОМУ ОТРЕЗАНИЮ

Малые группы с высоким весом отсеиваются, потому что они снижают результат.

CH08-F34— ПОДСЧЕТ ОЦЕНОК, ОСНОВАННЫХ НА МОДЕЛИ

Стране без примера присваивается определенная оценка.

CH08-F35— РАСКРЫТИЕ ЛИЧНОСТИ ДЛЯ ДОКАЗАТЕЛЬСТВ

Скриншоты публикуются с личной информацией.

CH08-F36— СДЕЛАТЬ УЧАСТНИКА ИЗ МАЛОЙ СТРАНЫ ПЕРЕИЗВЕСТНЫМ

Информация о стране, времени, устройстве и плане позволяет идентифицировать одного человека.

CH08-F37— ЗАПРОС ПАРОЛЯ ИЛИ ЛИЧНОГО ЧАТА У УЧАСТНИКА

Сбор доказательств нарушает безопасность аккаунта.

CH08-F38— ОТНОСИТЬ НИЗКИЕ ЦИФРОВЫЕ НАВЫКИ К НЕПОДХОДЯЩЕСТВУ

Фактические целевые пользователи удаляются из панели.

CH08-F39— НАРУШЕНИЕ ДОБРОВОЛЬНОСТИ С ПОМОЩЬЮ НАГРАДЫ

Участник находится под давлением, от которого он разумно не может отказаться.

CH08-F40— ОСТАВЬТЕ ВЫБОР УЧАСТНИКА ЗА КЛИЕНТОМ

Аудируемая организация выбирает людей, которые могут определять результат панели.

53. ПРОЦЕДУРА АУДИТА

Шаг 1— Карта источников участников

Каждая панель записывается отдельно для партнера, открытого набора и источника клиента.

Шаг 2— Назначение статуса выбора

Для каждого источника определяется соответствующий статус между PSI-0 и PSI-5.

Шаг 3— Сравните охват с целевой аудиторией

Определите, какие целевые пользователи отсутствуют на панели.

Шаг 4— Сохранение вероятности выбора участника или модели

Известная вероятность, квота или добровольный выбор объясняется.

Шаг 5— Заморозить правила соответствия

Возраст, страна, язык, продукт, план, интерфейс, уникальность, условия извлечения отношений версионированы.

Шаг 6— Просмотреть текст приглашения

Оценивается, присутствует ли положительное или отрицательное руководство.

Шаг 7— Изучить структуру вознаграждений

Подтверждается, что вознаграждение полностью независимо от результата.

Шаг 8— Проверить страну участника, язык и доступ к продукту

Назначаются уровни проверки с защитой конфиденциальности.

Шаг 9— Выполнить проверку уникальности человека

Проверяются повторения внутри источника и между источниками.

Шаг 10— Записать области отношений и знакомости

Определяется знакомство с клиентами, сотрудниками, конкурентами, поставщиками ИИ и брендами.

Шаг 11— Распределение задач согласно плану назначения

Проверяются назначения продукта, запроса, очереди и волны.

Шаг 12— Регистрация кондиционирования панели

Новые, повторяющиеся и ранее участвовавшие пользователи в одном и том же тесте разделяются.

Шаг 13— Настройка воронки участия

Подсчитываются приглашения, согласия, начало, завершение и показатели валидности.

Шаг 14— Сокрытие проверки валидности

Валидность протокола определяется без просмотра оценки AI-ответа настолько, насколько это возможно.

Шаг 15— Проверка сигналов мошенничества

Используются несколько сигналов и человеческая оценка.

Шаг 16— Анализ моего не ответа и потерь панели

Сравниваются известные характеристики участников и тех, кто не участвует.

Шаг 17— Генерация проектных и калибровочных весов

Вероятностные и неп 概率 стические источники обрабатываются отдельными методами.

Шаг 18— Расчет диагностик веса

Рассчитаны распределение веса, эффект дизайна, эффективная выборка, эффект обрезки.

Шаг 19— Выполнение анализа чувствительности

Сравниваются исходные, основные значительные и альтернативные разумные результаты.

Шаг 20— Создание манифеста конфиденциальности и доказательств

Полные доказательства находятся в защищенном хранилище; анонимные идентификаторы доказательств сохраняются в публичной записи.

Шаг 21— Публикация публичной избирательной карточки

Источники панели, воронка участия, смещения и эффект взвешивания становятся видимыми.

54. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Запись целевой аудитории; рамки выборки; поставщики панелей; источники участников; статус отбора; класс панели с вероятностной или непредвзятой выборкой; метод регистрации в панели; текст промпта; правила допуска участников; проверка страны участника; запись о владении языком; запись о доступе к продукту ИИ; план и пользовательский интерфейс;

проверка на уникальность личности; проверка дублирования по различным источникам; знакомство с брендом; частота использования ИИ; связь с проверяемой организацией; связь с поставщиком ИИ; компенсация участнику; условия бонусов; метод назначения задач; запись о рандомизации и блокировке; порядок продуктов и промптов; подготовка панели; участие в предыдущем пилотном исследовании; количество приглашённых; количество согласившихся; количество начавших задачу; количество завершивших задачу; количество действительных наблюдений; причины отсутствия ответа; отток панели; причины недействительности; и индикаторы мошенничества.

Записи обзоров и возражений Дизайнерские веса Настройки отсутствующих ответов Переменные калибровки Модель склонности, если есть Запись об усечении весов Эффективный размер выборки Результаты чувствительности Эффект поставщика панели Невзвешенные и взвешенные показатели Запись согласия Политика хранения и доступа к данным Правила редактирования Публичный реестр доказательств Ответственное лицо или организация

55. ЧЕК-ЛИСТ АУДИТА

Из каких источников пришли участники? Был ли каждый источник вероятностным, добровольным или открытым призывом? Была ли случайность внутри панели представлена как случайность мировой популяции? Вмешивалась ли контролирующая организация в выбор участников? Являются ли участники-клиенты и участники-сотрудники отдельными? Были ли задокументированы взаимоотношения с поставщиком ИИ или конкурентами? Как проверялась допустимость страны участника? Соответствует ли уровень владения языком фактическому использованию? Использовался ли правильный продукт ИИ, план и интерфейс? Сколько учетных записей или панелей использовал один и тот же человек? Проводилась ли перекрестная проверка между источниками? Влиял ли текст промпта на результат? Зависит ли награда от содержания ответа ИИ? Снизили ли стимулы за скорость качество протокола? Было ли назначение задач случайным или основанным на заранее заданных правилах? Выбирал ли участник продукт или задавал себе запрос самостоятельно?

Был ли продукт и порядок подачи сбалансированным? Было ли зарегистрировано кондиционирование панели? Делились ли участники результатами заданий? Видна ли воронка промпт–принятие–начало–завершение–валидность? Отличаются ли неоткликнувшиеся участники по известным характеристикам? Влияет ли отток панели на временной тренд? Было ли решение о валидности принято до оценки ответа ИИ? Считались ли отрицательные или критические ответы чаще недействительными? Несет ли обнаружение мошенничества несколько сигналов? Вызывали ли VPN или общие устройства автоматическое исключение? Был ли предоставлен участнику путь для апелляции? Были ли переменные весов определены до получения результата? Были ли неправдоподобные веса панели корректно маркированы? Надежны ли целевые итоги? Результаты без взвешивания и с взвешиванием представлены вместе? Ясна ли эффективная выборка? Показано ли влияние обрезки весов?

Показаны ли группы, которые никогда не наблюдались, как преимущественно представленные? Были ли предсказания на основе моделей представлены так, как если бы это были прямые наблюдения? Есть ли личные данные на скриншотах участников? Можно ли повторно идентифицировать пользователей из маленьких стран? Запрашивались ли пароли от аккаунтов или приватные чаты? Является ли согласие участников ясным и понятным? Есть ли явно ответственный владелец за выбор участников и учет их записей?

56. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Если мы используем профессиональную исследовательскую панель, почему мы все еще говорим о предвзятости?»

Профессиональная панель может повысить качество данных и управление на местах. Члены панели также:

- добровольный
- привыкший к исследованиям,
- различающийся в использовании технологий

Это возможно. Профессионализм не устраняет предвзятость. Он делает измерение и управление легче.

Возражение 2— «Если соблюдаются квоты по возрасту и стране, почему панель не может считаться репрезентативной?»

Участники:

- Частота использования ИИ,
- устройство,
- план,
- осведомленность о бренде,
- цифровые навыки

Это может различаться с точки зрения обслуживания. Квоты балансируют только те области, которые они охватывают.

Возражение 3— «Разве взвешивание не может скорректировать эти различия?»

Оно может частично скорректировать различия, для которых существуют измеренные и надежные целевые показатели. Оно не может окончательно исправить различия, которые никогда не наблюдаются или не измеряются.

Возражение 4— «Клиенты учреждения — это реальные пользователи; почему их не следует включать в основную панель?»

Их можно включить. Однако их следует включать с фактической долей и статусом отношений, которые они имеют в общей популяции. Весь список клиентов не является панелью общей популяции.

Возражение 5— «Почему это будет проблемой, если участник знает бренд?»

Это не обязательно должно быть проблемой. Они являются частью естественной пользовательской популяции. Если знакомство неравномерно, результат может варьироваться. Поэтому это фиксируется и анализируется при необходимости.

Возражение 6— «Пользователи с интенсивным использованием ИИ выполняют задачу более точно; не лучше ли их выбрать?»

Это может быть проще для тестирования протокола. Однако, если мы хотим измерить реальный опыт пользователей, неопытные пользователи также важны. Если задача слишком сложная, инструмент измерения должен быть исправлен. Популяция не должна быть упрощена.

Возражение 7— «Если участник неправильно понял ответ, почему мы должны сохранять это наблюдение?»

Задача участника не оценивать ответ, а правильно зафиксировать его. Если протокол фиксации применяется правильно, ответ ИИ сохраняется. Оценка материала является обязанностью судей.

Возражение 8— «Разве не следует использовать строгий автоматический фильтр, чтобы предотвратить фальшивое участие?»

Автоматическая фильтрация полезна. Если она становится единственным окончательным решением, это может несправедливо исключать реальных пользователей. Необходимы множественные сигналы и пропорциональный человеческий обзор.

Возражение 9— 'Если мы не опубликуем тысячу скриншотов, можно ли считать метод доказанным?'

Изображения на полный экран:

- личная информация,
- безопасность аккаунта,
- переопределение

может нести риск. Для публики:

- отредактированные примеры,
- манифест наблюдений,
- контрольные суммы целостности,
- доступ к независимому аудиту

может быть предоставлен. Доказуемость не равна раскрытию личности.

Возражение 10— “Если взвешенный балл сильно отличается от исходного, какой из них правильный?”

Два результата отвечают на разные вопросы. Сырой балл — это результат участников панели. Взвешенный балл — это оценка целевой популяции. Большая разница указывает на то, что состав панели имеет значение. Оба значения следует публиковать.

Возражение 11— «Если эффективная выборка сократилась до 349, было ли сбор 1,000 наблюдений напрасным?»

Переизбыточно выбранные группы могли усилить анализ подгрупп. Однако чувствительность глобального прогноза может быть ниже, чем сырое количество 1,000. Этот факт не следует скрывать.

Возражение 12— «Можно ли вообще не использовать панель без вероятностного отбора?»

Можно использовать. В частности:

- пилот,
- язык с ограниченными ресурсами,
- маленькая страна,
- Быстрое исследование

Это ценно для. Его следует публиковать с правильным статусом и ограничениями.

Возражение 13— «Разве не надежнее выбирать всех участников из новых людей?»

Это может быть сильным показателем для поперечной оценки населения. Повторные пользователи ценны для измерения продольных изменений. Новые и повторные участники панели должны оставаться отдельными.

Возражение 14— «Разве сбор данных у такого большого числа участников не увеличивает риск для конфиденциальности?»

Может. Поэтому:

- минимизация данных,
- анонимная идентификация,
- редактирование,
- контроль доступа,
- срок хранения

являются обязательными частями протокола. Ненужная информация не должна собираться.

ОБЩАЯ ПОЛОЖЕНИЕ РАЗДЕЛА 59

Самое простое число в GEO-исследовании:

1,000

Это возможно. Самый трудный вопрос таков:

> *Какой 1,000?*

Две панели с одинаковыми странами и возрастными квотами могут дать совершенно разные результаты. В одной:

- активные пользователи ИИ,
- подписчики платного плана,
- работники в сфере технологий

составляют большинство. В другой:

- редко использующие,
- мобильные пользователи,
- люди с низкими цифровыми навыками

Это более заметно. Оба человека являются 1,000. Это не одно и то же население. Если вы скрываете, откуда участники, и публикуете только взвешенный результат, вы превращаете

взвешивание в занавес доверия, а не в метод контроля. Вы не можете взвесить человека, который не присутствует в выборке. Вы не можете изменить ответ пользователя только потому, что он неправильный. Неправильный ответ — это не провал исследования. Это событие, которое исследование пытается измерить. Если участник неправильно следовал протоколу, несмотря на правильный ответ, вы не можете сохранить запись. Правильный ответ не делает ложное наблюдение истинным. Вы можете платить одному человеку больше. Но не для того, чтобы получить нужный вам ответ. Вы можете платить в соответствии с условиями жизни в стране. Но вы не можете купить готовность участника. Тысяча скриншотов может служить доказательством. Вы не обязаны раскрывать личность тысячи человек публике.

Целостность доказательств и конфиденциальность личности могут быть защищены вместе. Следовательно, восьмой закон измерений NOMOS следующий:

> *Мощность выборки определяется не только количеством людей, которые в нее входят, но и тем, почему и как были выбраны эти люди.*

Девятый закон звучит следующим образом:

> *Взвешивание уравновешивает наблюдаемую популяцию; оно не создает ненаблюдаемого человека.*

Десятый закон таков:

> *Некорректный ответ III может быть действительными данными; правильный ответ III может быть недействительными данными.*

Одиннадцатый закон звучит следующим образом:

> *Награду участнику следует давать не за нахождение желаемого результата, а за доказательство наблюдаемого результата без его изменения.*

Двенадцатый закон звучит следующим образом:

> *Доказуемость не требует от участника отказа от приватности.*

Заповедь главы NOMOS 8

> *Не приводите ко мне тысячу людей в одиночку. / Покажите мне, из какого мира пришли эти тысяча людей.*
>
> *Не рассматривайте панель добровольцев как случайную выборку населения, список клиентов как общественность или сотрудника как независимого пользователя.*
>
> *Не думайте, что все различия, которые вы не измеряли, улучшились только потому, что были соблюдены квоты по стране и возрасту.*
>
> *Не умножайте одного человека на количество его аккаунтов, устройств, языков или участий в панелях.*
>
> *Не помещайте тех, кто любит бренд, тех, кто ненавидит бренд, и тех, кто вообще не знает бренд, в одну невидимую коробку.*
>
> *Платите участнику не за то, чтобы найти правильный ответ, а за то, чтобы сохранить первый подходящий ответ, который они увидели, не меняя его.*
>
> *Не меняйте пользователя, когда III дал неправильный ответ. / Не проверяйте правильность ответа, когда пользователь нарушает протокол.*
>
> *Не заставляйте человека, который не состоит в панели, появляться с весом, как будто он существует.*

Не скрывайте необработанный результат. / Не объявляйте истиной только взвешенный результат. / Показывайте эффективную выборку. / Показывайте её чувствительность.

Не объявляйте кого-либо использующего VPN поддельным, кого-либо использующего общий девайс - ботом, или кого-либо использующего вспомогательные технологии - автоматизированной системой.

Расследуйте мошенничество. / Но не превращайте само расследование в новый предвзятый отказ.

Сделайте скриншот. / Защитите имя человека, аккаунт и личную жизнь.

Сначала определите популяцию. / Затем покажите структуру. / Затем объясните, кого вы выбрали, кто не пришёл, кто не смог завершить, и почему кого-то исключили. / И только после этого решите, насколько ваш результат может говорить о человечестве.

Заключительное предложение главы

В GEO-1000 участник — это не просто человек, предоставляющий ответ. Участник — это человеческая единица выборки, способность которой представлять других в целевой популяции должна быть обоснована через путь выбора, вес, независимость и меры защиты приватности.

Нормативное ядро

Каждое наблюдение GEO-1000 ДОЛЖНО сохранять поддающийся аудиту путь выбора участников от целевой популяции и выборки до этапов отбора, приглашения, согласия, распределения, завершения, проверки протокола, исключения и окончательного взвешивания. Случайный отбор внутри панели НЕ ДОЛЖЕН представляться как случайный отбор из глобальной целевой популяции, если механизм включения всей популяции не поддерживает это утверждение. Набор участников, распределение, компенсация, проверка, замена, исключение и взвешивание ДОЛЖНЫ оставаться независимыми от того, является ли ответ ИИ положительным, отрицательным, правильным, неправильным, отклонённым, критическим или коммерчески благоприятным. Лица, аккаунты, устройства, сессии, языки, членства в панелях и наблюдения за продуктами ИИ ДОЛЖНЫ оставаться отдельными единицами. Источники участников с вероятностью участия, с добровольным присоединением, с открытым приглашением, предоставленные клиентом, институциональные и гибридные ДОЛЖНЫ быть идентифицированы отдельно. Взвешивание МОЖЕТ уменьшать наблюдаемый дисбаланс, но НЕ ДОЛЖНО представляться как создание наблюдений для популяций, отсутствующих в выборочной рамке, или как исправление неизмеренного отбора с уверенностью. Невзвешенные результаты, взвешенные результаты, диагностика весов, пропуски покрытия, неответы, эффективный размер выборки и анализ чувствительности ДОЛЖНЫ оставаться видимыми. Решения о мошенничестве и дублировании ДОЛЖНЫ использовать пропорциональный обзор с несколькими сигналами и НЕ ДОЛЖНЫ автоматически рассматривать использование VPN, совместные устройства, корпоративные сети или средства доступности как подделку. Доказательства участников ДОЛЖНЫ оставаться проверяемыми без излишнего раскрытия личности, учетных данных, частных разговоров или точного местоположения. Каждое решение о выборе участников и взвешивании ДОЛЖНО иметь версию и быть приписано ответственному человеку или организации.

ЧАСТЬ III

Проведение эксперимента и цепочка доказательств

ГЛАВА 09

Реестр ИИ-систем и сопоставимость

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K02; K03

Граница главы

Первые восемь глав установили следующее: один ответ не является оценкой GEO; представление является распределением; проверяемый объект должен быть идентифицирован до наблюдения результатов; официальное представление и проверенная реальность объекта должны оставаться раздельными; для каждого продукта ИИ необходимо определить допустимую пользовательскую популяцию; действительные наблюдения 1,000 должны распределяться с использованием выборочного дизайна на основе популяции; меньшие страны и языки должны оставаться видимыми через дополнительные панели; а выбор участников, их исключение и взвешивание должны оставаться независимыми от содержания ответа. Теперь мы должны определить другую сторону эксперимента:

> *Какую ИИ-систему участники фактически измеряют?*

Название провайдера? Базовая модель? Чат-продукт? Бесплатный веб-интерфейс? Платное мобильное приложение? Сеанс с включенным веб-поиском? Режим ответа без ссылок на источники? Корпоративное рабочее пространство, поддерживаемое внутренними данными? Аккаунт с включенной памятью? Временный или очищенный чат? Являются ли разные пользовательские интерфейсы одного и того же провайдера одной и той же системой? Следует ли измерять два продукта, использующие одну и ту же семейство моделей, одинаковым образом? Даже если продукт отображает одно и то же название модели на экране:

- получение,
- выбор источника,
- правила безопасности,
- персонализация,
- доступ к инструментам,
- экспериментальные функции

Если это изменяется, можно ли сравнивать результаты? Когда ИИ-продукт обновляется в середине измерительной волны, можно ли считать первые и последние ответы волны исходящими от одной и той же системы? Пользователь:

- веб-поиск включен,
- память включена,
- имеет предыдущие разговоры,
- платный тарифный план

ответы; другого пользователя:

- поиск выключен,
- память выключена,
- новая сессия,
- бесплатный план

Можно ли объединить это в одной ячейке с его ответом? Нет. Принцип измерений, установленный предыдущими разделами, ясен: нет единого «ИИ»; индекс поиска, система извлечения, переоценщик, языковая модель, слой цитирования и система безопасности могут вести себя по-разному. Одна и та же структура требует, чтобы каждый проверяемый результат содержал:

- модель или продукт,
- дата,
- страны,
- языка,
- набор запросов,
- число повторений,
- запись измерения

Этот раздел применяет эти принципы в Реестре систем ИИ. Этот раздел:

- различие между поставщиком, продуктом, моделью и пользовательский интерфейс,
- экземпляр продукта ИИ,
- уровни уверенности идентичности модели,
- конфигурации поиска и инструментов,
- состояния памяти и персонализации,
- отличия в плане, интерфейсе и экспериментальных функциях,
- изменения в продукте,
- конфигурация системы на момент измерения,
- уровни сопоставимости между продуктами,
- сравнения Естественный охват и Общая опора сравнения,
- дрейф версии со временем,

определяет. В этой главе пока не рассматривается:

- система полной транслитерации и семантической эквивалентности промптов,
- все детали контролируемых и естественных пользовательских панелей,
- архитектура программного обеспечения захвата NOMOS,
- атомная оценка утверждений ответов,
- финальные коэффициенты модели и формула оценки NOMOS

не завершена. Ключевой вопрос Раздела 9 заключается в:

> *Прежде чем сравнивать два ответа на один и тот же запрос, как мы докажем, что эти ответы действительно исходят из сопоставимых условий работы ИИ-продукта?*

NOMOS Вызов

Два пользователя одного и того же провайдера задают один и тот же вопрос: «Что это за компания Apple.com?» Первый пользователь:

- использует бесплатный веб-продукт,
- открывает новый чат,
- память выключена,
- не имеет доступа к веб-сайтам,

получает ответ только с использованием встроенной в продукт возможности знаний. Второй пользователь:

- Использование платного мобильного продукта,
- память включена,
- ранее говорил об Apple,
- веб-поиск включен,

получает ответы с актуальными источниками и цитатами. Первый ответ: «Apple — это технологическая компания, которая производит потребительскую электронику.» Второй ответ: «Apple — глобальная технологическая компания, предлагающая аппаратное обеспечение, программное обеспечение и цифровые услуги. Актуальную корпоративную информацию можно увидеть в этих источниках.» Оба ответа исходят от одного и того же поставщика.

Это результат одной и той же системы? Нет. Теперь рассмотрим двух других пользователей. Оба видят на экране одно и то же название модели. В стране одного пользователя веб-функция включена. В стране другого пользователя она выключена. В ответе одного пользователя видны исходные ссылки. Интерфейс другого пользователя не показывает цитаты, даже если использует те же источники. Модель имеет одинаковую метку.

Есть ли та же измерительная ячейка? Нет. Теперь рассмотрим два продукта от разных поставщиков. Оба:

- новый чат,
- память выключена,
- один и тот же язык,
- та же страна,
- один и тот же временной интервал,
- веб-доступ включен,
- одна и та же версия промпта

измеряются в этих условиях. Они основаны на разных базовых моделях. Можно ли их сравнивать? Да, их можно сравнивать с точки зрения конкретного результата для пользователя. Потому что объект сравнения:

> *не в том, используют ли они одну и ту же базовую модель, а в том, какое представление они создают при одинаковых определённых условиях для пользователя.*

Теперь пусть версия продукта остаётся неизменной для первых 500 наблюдений за измерительной волной. В течение следующих 500 наблюдений поставщик:

- изменяет систему извлечения,
- правила безопасности,
- интерфейс пользователя

Сам отчёт говорит: «Пользователи 1,000 тестировали один и тот же продукт ИИ». На самом деле могли быть измерены две разные продуктовые ситуации. Первое положение этого раздела следующее:

> *Имя поставщика недостаточно для идентификации измеряемой системы ИИ.*

Второе положение гласит:

> *Название модели не отражает полностью поведение продукта, которое достигает пользователя.*

Её третье положение гласит:

> *Один и тот же поставщик, одна и та же модель или один и тот же логотип не гарантируют сопоставимости.*

Четвертая статья гласит:

> *Разные поставщики могут быть сопоставимы с точки зрения результатов для пользователей, если условия измерения достаточно совпадают.*

Пятая его статья гласит:

> *Если продукт существенно меняется во время измерения, его нельзя представлять как одну фиксированную систему.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — определить ИИ-продукт, проверенный в GEO-1000 со всеми его материальными конфигурациями, и определить, при каких условиях можно сравнивать различные результаты продукта. Раздел стандартизирует следующие различия:

- Поставщик и ИИ-продукт
- ИИ-продукт и базовая модель
- Семейство моделей и конкретное развертывание
- Метка модели и проверенная идентичность модели
- пользовательский интерфейс и API
- Веб- и мобильный продукт
- Бесплатный план и платный план
- Потребительский продукт и корпоративный продукт
- Базовый ответ модели и ответ с расширенным поиском
- Прямой веб-поиск с использованием поиска
- Отображение цитат с указанием источника
- Использование инструмента и доступ к инструменту
- Контекст сессии и память
- Системная политика с пользовательской инструкцией
- Материальная реальность с персонализацией
- Временный чат с новым разговором
- Статус продукта с течением времени с фиксированным названием продукта
- Наблюдаемое изменение поведения с обновлением, объявленным провайдером
- Сравнение Естественный охват и Общая опора сравнения
- Пример сопоставимого продукта в рамках одной продуктовой линейки
- Результат пользователя versus абстрактная способность модели
- Отсутствие продукта versus сбой продукта
- Техническая ошибка versus фактическое отклонение
- Одинаковый запрос versus тот же эксперимент

- Одинаковое среднее versus одинаковое поведение системы
- Мгновенный снимок продукта versus долгосрочная производительность продукта

В конце этого раздела каждая проверка GEO-1000 должна быть способна ответить на следующие вопросы:

> *Какого провайдера, какой продукт, какая интерфейс, какой план и какая конфигурация были измерены?*
>
> *Насколько действительно была проверена модель с указанным на экране именем?*
>
> *Каковы были статус веба, поиска, цитирования, памяти, персонализации и инструментов?*
>
> *Оставался ли продукт неизменным на протяжении всей проверки?*
>
> *На каком уровне сопоставимости два результата продукта были размещены рядом?*
>
> *С чем сравнивается объект: с базовой моделью, потребительским продуктом, пользовательский интерфейс или всей системой цепочки?*

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

> *Каждое наблюдение GEO-1000 должно быть связано не только с поставщиком или названием модели; оно должно быть связано с записью в версионированном Реестре AI-систем, идентифицирующей экземпляр ИИ-продукта на момент измерения.*

Экземпляр ИИ-продукта должен содержать как минимум следующие поля, насколько они релевантны:

- Поставщик
- Название продукта
- пользовательский интерфейс
- Тип продукта
- План или класс учетной записи
- Видимая метка модели
- Уровень уверенности в идентификаторе модели
- Версия продукта или развертывания
- Страна и локализация
- Язык
- Веб- или поисковый доступ
- Функция цитирования
- Доступ к инструментам
- Статус памяти
- Статус пользовательских инструкций
- Контекст предыдущего чата
- Тип сессии
- Статус персонализации
- Режим безопасности или политики, если известен
- Экспериментальные функции
- Дата измерения и время UTC
- Измерительная волна

- Журнал изменений продукта
- Достоверность доказательств и записей

Когда любое из этих полей изменяется существенно:

- новый экземпляр ИИ-продукта,
- новая экспериментальная ячейка,
- новая версия анализа

может потребоваться.

3. «СИСТЕМА ИИ» НЕ ЯВЛЯЕТСЯ ОДНИМ ОБЪЕКТОМ [K02; K03]

В этом протоколе термин «система ИИ» является общим и высокоуровневым. Реальный ответ пользователя может возникнуть в результате комбинации нескольких слоев. Кандидатская цепочка:

```
R= F(M, O, K, Q, T, P, S, U, C, L, t)
```

Здесь:

- M: базовая или направляющая модель
- O: слой оркестровки продукта
- K: слой поиска или источников знаний
- Q: перенарядка и выбор источника
- T: инструменты
- P: слой политики и безопасности
- S: сессия, память и персонализация
- U: учетная запись пользователя и план
- C: страна и локаль
- L: язык
- t: время

Ответ, видимый пользователю:

R

Ответ, видимый пользователю, не обязательно является результатом работы только базовой модели. Основной объект, ориентированный на потребителя, измеряемый GEO-1000, поэтому:

> *поведение продукта, представленное пользователю, а не абстракция модели.*

Можно также проводить контролируемые исследования API или моделей. Это не тот же результат, что и основная панель потребителей.

4. СЕМЬ СЛОЕВ ИДЕНТИЧНОСТИ СИСТЕМЫ ИИ

Каждая проверяемая система ИИ должна быть определена по семи слоям, насколько каждый слой релевантен.

4.1. Слой поставщика

Организация, которая публикует или управляет продуктом. Название поставщика:

- не то же самое, что

  ID продукта,
- ID модели,
- пользовательский интерфейс

Поставщик может предлагать более одного продукта. В рамках одного продукта может быть использовано более одной модели или системы оркестровки.

4.2. Слой продукта

Коммерческий, потребительский, корпоративный или технический продукт, к которому получает доступ пользователь. Примеры типов:

- Общий чат-продукт
- Продукт с поисковой поддержкой для ответов
- Корпоративное рабочее пространство
- Образовательный продукт
- Встроенный помощник
- API для разработчиков
- Продукт с пользовательской базой знаний

Название продукта само по себе не описывает всю конфигурацию.

4.3. Уровень пользовательский интерфейс

Это интерфейс, через который пользователь взаимодействует с продуктом. Веб, Мобильный, Настольный, Расширение браузера, Интеграция с операционной системой, Встроенный интерфейс в платформе

API

Корпоративный портал Разные интерфейсы одного и того же продукта:

- разные функции,
- разная модель,
- разная форма представления источника,
- разная память

может использоваться.

4.4. Модель или уровень инференса

Это модель, семейство моделей или система маршрутизации, используемая при генерации ответа. ID модели:

- открытая,
- частично видимая,
- неизвестно,
- динамическая

Это возможно. Продукт может направлять запрос пользователя к разным моделям. В этом случае нельзя сказать: «Была измерена одна фиксированная модель».

4.5. Слой оркестрации и информации

Он может включать следующее:

- Извлечение
- Поиск в живой сети
- Внутренний поиск
- Ранжирование источников
- Переранжирование
- Выбор цитат
- Вызов транспортного средства
- Переписывание запросов
- Измельчение документов
- Объединение ответов

Этот слой имеет фундаментальное значение для представления GEO. Поскольку он может влиять на то, какая запись:

- будет получена,
- будет выбрана,
- будет включена в ответ,
- будет показана как ссылка

может быть подвержена влиянию.

4.6. Слой политики и персонализации

Он может включать следующее:

- Правила безопасности
- Системные инструкции
- Политика продукта
- Память
- Особые инструкции
- История пользователя
- Региональные правила
- Возраст или класс учетной записи
- Настройки корпоративного администратора

Этот слой влияет на:

- объем ответа,
- его тон,
- его отклонение,
- его рекомендации

может изменять.

4.7. Распределение и временной слой

Указывает, к какому из следующих состояний продукт относится в данный момент:

- имеет смысл с
- экспериментальная функция,

- региональное распределение,
- сегмент пользователей,
- A/B-тест

]

Одинаковое наименование продукта может иметь разные условия распределения для различных пользователей в один и тот же день.

5. ПРИМЕР AI ПРОДУКТА

Основной системный блок внутри GEO-1000:

Экземпляр AI продукта

Канонический операционный термин:

Экземпляр AI продукта

Экземпляр AI продукта определяется как:

```
A_i= (v, p, s, r, m, k, t, x, u, c, l, ω)
```

может быть представлен следующим образом. Здесь:

- v: поставщик
- p: продукт
- s: пользовательский интерфейс
- r: план или класс учетной записи
- m: ID модели или метка
- k: извлечение и структурирование информации
- t: конфигурация инструмента
- x: память, персонализация и состояние сеанса
- u: класс настроек пользователя или организации
- c: страна и местные настройки
- l: язык
- ω: время, развертывание и состояние экспериментальных функций

Наблюдение: Oj должен быть связан с конкретным: образцом продукта Ai.

6. РАЗЛИЧИЕ МЕЖДУ ОБРАЗЦОМ ПРОДУКТА И НАЗВАНИЕМ ПРОДУКТА

Название продукта может быть фиксированным. Образец продукта может различаться. Пример: Пусть название продукта Orion AI останется без изменений. Однако эти два измерения являются разными образцами продукта:

Образец продукта 1

Веб-план Free, память выключена, доступ в веб отключен, новый чат, турецкий, Турция, Wave 1

Образец продукта 2

Мобильный платный план Память на веб-доступе Доступен на предыдущем контексте чата Турецкий Турция Wave 1 Эти могут существовать под одним и тем же брендом. Они не являются одной и той же экспериментальной ячейкой.

7. УРОВНИ УВЕРЕННОСТИ ИДЕНТИЧНОСТИ МОДЕЛИ

Название модели не всегда можно узнать с одинаковой уверенностью. Поэтому идентичность модели может записываться на следующих уровнях.

MI-0— НЕИЗВЕСТНАЯ МОДЕЛЬ

Неизвестно, какую модель или семейство моделей использует продукт. Результат приписывается только интерфейс продукта.

MI-1— МАРКИРОВКА, ПОКАЗАННАЯ ПОЛЬЗОВАТЕЛЮ

Маркировка модели видна в пользовательском интерфейсе. Маркировка фиксируется. Не предполагается, что она полностью подтверждает развертывание на серверной части.

MI-2— СЕМЕЙСТВО МОДЕЛЕЙ, ДОКУМЕНТИРОВАННОЕ ПРОВАЙДЕРОМ

Документы продукта поставщика описывают конкретное семейство моделей или отношение продукта к модели. Все еще может быть неизвестно, направлено ли отдельное наблюдение именно на эту модель.

MI-3— РАЗВЕРТЫВАНИЕ С ПОДТВЕРЖДЕННОЙ СЕССИЕЙ

Сессия проверяет конкретное распределение API или доверенной метаданных продукта.

MI-4— ВЕРСИОНИРОВАННАЯ ТОЧКА ДОСТУПА ИЛИ РАЗВЕРТЫВАНИЕ

Можно записать конкретную версию, идентификатор точки доступа или развёртывания. Это более надёжно для контролируемого сравнения.

MI-5— ВОСПРОИЗВОДИМЫЙ ЗАМОРОЖЕННЫЙ АРТЕФАКТ МОДЕЛИ

Существует статическая и воспроизводимая структура модели или развёртывания. Это может быть редко в потребительских продуктах в реальном времени. Если доверие к идентификатору модели низкое, результат:

следует приписывать поведению продукта

Невозможно определённо приписать конкретной модели.

8. ПРЕДЕЛ ВИДИМОГО ИМЕНИ МОДЕЛИ

Видеть имя модели в пользовательском интерфейсе полезно. Однако следующие вопросы могут оставаться открытыми: Все ли запросы направляются к одной и той же модели? Меняется ли модель в зависимости от сложности запроса? Использует ли безопасность или поиск отдельную модель? Применяет ли поставщик одинаковую метку к различным распространениям? Указывает ли метка модели только на семейство продуктов? Является ли пользователь частью экспериментального развертывания? Следовательно, правильный отчет может выглядеть так: «Метка модели X была показана в пользовательском интерфейсе». Без

более веских доказательств не следует использовать следующее предложение: «Все ответы определенно генерируются исключительно моделью X.»

9. ТИП ПРОДУКТА

Каждый продукт ИИ должен быть классифицирован с основным типом и второстепенными типами.

PT-1— Общий разговорный продукт

Это общий чат и продукт для вопросов и ответов.

PT-2— Продукт, основанный на поиске или извлечении информации

Поиск или извлечение информации является фундаментальной частью генерации ответов.

PT-3— Продукт корпоративных знаний

Он может использовать внутренние документы, рабочие пространства или специальные источники данных.

PT-4— Встроенный помощник

Он встроен в другой продукт или платформу.

PT-5— Продукт для разработчиков или с API

Он используется разработчиком или приложением вместо конечного пользователя.

PT-6— Продукт для конкретной области

Он предназначен для конкретной профессии или отрасли.

PT-7— Мультимодальный продукт

Входные данные, отличные от текста, такие как визуальные, аудио или документальные записи, являются частью основного опыта. Продукт может включать более одного типа. Сравнение следует проводить в пределах одного типа или с явной областью.

10. РЕГИСТР пользовательский интерфейс

Разные интерфейсы одного и того же продукта должны записываться отдельно. Площади интерфейсы:

- ID интерфейсы
- Веб/мобильное/настольное/API
- Версия приложения
- Операционная система, если применимо
- Браузер, если применимо
- Язык продукта
- Видимое выбор моделей
- Доступность инструментов
- Просмотр цитирования

- Функция загрузки файлов
- Настройки памяти
- Настройки корпоративного управления
- Экспериментальные функции
- Время измерения

Если интерфейсы веб и мобильного приложения являются существенно одинаковыми, результаты могут быть объединены. Это равенство не следует предполагать; его необходимо протестировать.

11. ПЛАН И КЛАСС УЧЕТА

Тип плана может иметь функции, которые могут повлиять на ответ пользователя. Статусы плана:

- Бесплатно
- Платный индивидуальный
- Профессиональный
- Корпоративный
- Образование
- Для разработчиков
- Пробный
- По приглашению или с ограниченным доступом
- Неизвестно

Планы могут отличаться по следующим параметрам:

- Модель
- Лимит использования
- Доступ через веб
- Память
- Инструменты
- Ссылка на источник
- Длина ответа
- Скорость или приоритет
- Данные предприятия
- Настройки безопасности и хранения

Если информации о плане нет:

ПЛАН_НЕИЗВЕСТЕН

статус следует использовать. Бесплатные и платные результаты не могут быть автоматически объединены как одно и то же поведение продукта.

12. ДОСТУП И ИСПОЛЬЗОВАНИЕ ВЕБ

Ответ ИИ может быть сгенерирован в одной из трёх основных состояний знаний.

KG-0— РЕЖИМ НЕОПРЕДЕЛЁННЫХ ЗНАНИЙ

Невозможно определить случай использования извлечения или веб.

KG-1— ВНУТРЕННИЙ ОТВЕТ МОДЕЛИ

Ответ известен или с высокой степенью достоверности предполагается, что не использует живую или внешнюю информацию.

KG-2— ОТВЕТ С ПОДДЕРЖКОЙ ИЗВЛЕЧЕНИЯ

Могло быть использовано извлечение из предварительно индексированных или встроенных источников.

KG-3— ОТВЕТ С ПОДДЕРЖКОЙ ЖИВОГО ВЕБА ИЛИ ПОИСКА

Был использован доступ к веб/поиску в режиме реального времени или близко к времени измерения.

KG-4— ОТВЕТ, ОСНОВАННЫЙ НА ПРИВАТНЫХ ЗНАНИЯХ

Был использован приватный источник данных, предоставленный учреждением или пользователем.

KG-5— ГИБРИДНЫЙ ИЛИ МНОГОИСТОЧНИКОВЫЙ ОТВЕТ

Были использованы несколько слоев информации вместе. Эти режимы могут давать разные результаты для одного и того же запроса. Их следует чётко различать при сравнении.

13. ДОСТУП К ИЗВЛЕЧЕНИЮ НЕ ТОЖЕ, ЧТО И ИСПОЛЬЗОВАНИЕ ИЗВЛЕЧЕНИЯ

Продукт может иметь функцию веб-поиска. Однако она могла не использоваться в конкретном ответе. Три поля должны фиксироваться отдельно: была ли функция доступна? Была ли функция включена или выбираема? Было ли доказательство того, что она использовалась в конкретном ответе? Пример статусов:

ДОСТУПНО_НО_НЕ_ИСПОЛЬЗОВАНО

ДОСТУПНОСТЬ_ИСПОЛЬЗОВАНИЯ_НЕИЗВЕСТНА

ИСПОЛЬЗОВАНО_ПОДТВЕРЖДЕНО

НЕДОСТУПНО

ОТКЛЮЧЕНО

НЕИЗВЕСТНО

Само наличие ссылки на источник может быть сильным индикатором использования. Отсутствие ссылки на источник не обязательно доказывает, что извлечение не использовалось.

14. СЛОЙ ЦИТИРОВАНИЯ

Использование источника не то же самое, что показ цитаты. Продукт:

- может использовать источник, но может его не показывать,
- может показывать ссылку, но ссылка может не подтверждать утверждение,
- может связывать один и тот же источник с разными предложениями,

Пользователь может скрыть цитату в соответствии с пользовательский интерфейс. Запись цитаты должна содержать следующие поля: доступна ли функция цитирования? Была ли показана цитата в конкретном ответе? Количество цитат Названия полей источников Совпадение цитаты с утверждением Видимость на пользовательский интерфейс Расширяемость цитаты Доступность источника Время создания цитаты Продукты, показывающие и не показывающие цитаты, нельзя напрямую сравнивать по "количеству источников". Сначала необходимо объяснить различия в поведении продукта и интерфейсе.

15. ДОСТУП К ИНСТРУМЕНТУ

Продукт ИИ может получить доступ к следующим инструментам:

- Веб-поиск
- Браузер
- Калькулятор
- Выполнение кода
- Чтение файлов
- Карта
- Календарь
- Корпоративная база знаний
- Каталог продукции
- Другие API

Доступ к инструментам может повлиять на ваш ответ:

- его актуальность,
- доказательства,
- объем ответа,
- точность

Реестр инструментов:

- доступность инструмента,
- активен ли он,
- был ли он использован в ответе,
- сохраняется ли результат инструмента

должно быть показано.

16. ПАМЯТЬ, СПЕЦИАЛЬНЫЕ ИНСТРУКЦИИ И ПЕРСОНАЛИЗАЦИЯ

Ответ, видимый пользователю, может быть затронут следующими полями:

- Постоянная память
- История чата
- Особые инструкции
- Имя пользователя или профиль
- Предыдущие разговоры с брендом
- Предпочтения
- Страна и локализация
- Настройки корпоративного рабочего пространства
- Сегментация, выполненная поставщиком

Эти поля должны быть записаны в трех отдельных слоях.

16.1. Постоянная персонализация

Информация защищена во всех учетных записях или сессиях.

16.2. Контекст в разговоре

Предыдущие сообщения в текущем чате.

16.3. Системные или организационные настройки

Администратор или инструкции продукта, которые пользователь может не видеть. Полная информация может быть недоступна. Неизвестные ситуации должны быть явно зафиксированы.

17. ТИПЫ СЕССИЙ

Статусы кандидатов на сессию:

ST-1— НОВЫЙ СТАНДАРТНЫЙ РАЗГОВОР

Новый стандартный разговор. Память или специальные инструкции записываются отдельно.

ST-2— ВРЕМЕННЫЙ ИЛИ ЭФЕМЕРНЫЙ РАЗГОВОР

Временная сессия, предназначенная для отделения от постоянной истории. Фактические функции следует проверять в реестре продукта.

ST-3— СУЩЕСТВУЮЩИЙ РАЗГОВОР

Есть контекст предыдущего разговора. Он может быть ценен для естественной панели. Он хранится отдельно в контролируемой панели.

ST-4— СЕССИЯ ПРЕДПРИЯТИЯ ИЛИ ОБЩЕГО РАБОЧЕГО ПРОСТРАНСТВА

Может существовать область для корпоративной или общей информации.

ST-5— НЕИЗВЕСТНОЕ СОСТОЯНИЕ СЕССИИ

Статус сессии не удалось определить надёжно. Если разные типы сессий должны быть объединены в одной ячейке, необходимо объяснить обоснование и веса.

18. ПОЛИТИКА ПРОДУКТА И ОТКАЗ

Продукт ИИ может не отвечать на конкретный вопрос по следующим причинам:

- безопасность,
- закон,
- политика продукта,
- региональные правила,
- не подразумевает никакой пригодности в отношении возраста,
- настройки предприятия

причина. Отказ:

- базовые недостатки модели,
- политика продукта,
- интерфейс,
- может быть связан с учетной записью пользователя

GEO-1000 основные записи управления пользователем фиксируют результат отказа, доходящий до пользователя. Однако, если причина отказа известна, она дополнительно классифицируется. Слой политики продукта также является частью пользовательского опыта.

19. ЭКСПЕРИМЕНТАЛЬНЫЕ ФУНКЦИИ И A/B РАСПРЕДЕЛЕНИЯ

Продукты ИИ в реальном времени могут предоставлять не все функции всем пользователям одновременно. Участники:

- разный пользовательский интерфейс,
- различные формулировки промптов для разных моделей,
- разные веб-функции,
- разный формат ответа

может видеть. Индикаторы экспериментального статуса:

- Бета-метка на интерфейсе
- Документация поставщика
- Различие в функционале
- Систематическое различие интерфейса между пользователями
- Нарушение поведения во время измерения

может происходить. Если экспериментальная функция неизвестна:

EXPERIMENTAL_STATE_UNKNOWN

должна использоваться. Фактическое различие внутри продукта не является дефектом данных. Это особенность распределения системы. Однако её следует учитывать при сопоставимости.

20. Реестр ИИ-систем

Для каждого проверяемого продукта должен быть создан реестр ИИ-систем. Реестр должен иметь три уровня.

20.1. Основная запись поставщика и продукта

Включает относительно постоянные области:

- Поставщик
- Название продукта
- Тип продукта
- Основные пользовательские интерфейсы
- Страны с официальным доступом
- Классы планов
- Записи поддерживаемых языков
- Общие функции

20.2. Запись продукта Measurement Wave

Для конкретной волны:

- версия продукта,
- функции плана,
- метки модели,
- веб- и инструментальные функции,
- поведение памяти,
- статус доступа,
- известные обновления

записываются.

20.3. Запись системы на уровне наблюдений

Фактически наблюдалось для каждого участника:

- интерфейс,
- метка модели,
- план,
- использование транспортного средства,
- цитата,
- сессия,
- персонализация,
- время

несет поля. Основная запись сама по себе не может устранить различия на уровне наблюдений.

21. ОТПЕЧАТОК КОНФИГУРАЦИИ СИСТЕМЫ

Отпечаток конфигурации системы может быть создан для облегчения сопоставимости экземпляра продукта ИИ. Кандидатные поля:

```
SCF= H(canonicalise(p, s, r, m, k, t, x, c, l, w))
```

Здесь H — это функция, которая производит значение целостности из версионированных полей. Отпечаток:

- определяет, принадлежат ли два наблюдения одной и той же конфигурации,

- определяет, были ли существенные изменения в потоке

Это может помочь в идентификации. Отпечаток не доказывает реальную модель бэкенда, не делает видимыми слои, которые провайдер не раскрыл. Он лишь подтверждает, было ли изменено записанное состояние конфигурации позже.

22. ЧТО ТАКОЕ СОПОСТАВИМОСТЬ?

Два результата считаются сопоставимыми, когда:

> *Разница между ними может быть достаточно объяснена отслеживаемым поведением продукта, а не несущественными различиями в дизайне измерений.*

Сравнимость - это не абсолютное равенство. Разные продукты ИИ:

- модель,
- поставщик,
- пользовательский интерфейс.

Не обязательно должны быть идентичными. Однако материальные условия, относящиеся к сравнению, должны быть:

- выравнивание,
- многоуровневость,
- запись,
- моделирование в анализе

требуются.

23. ОБЪЕКТ СРАВНЕНИЯ ДОЛЖЕН БЫТЬ ОПРЕДЕЛЕН ЗАРАНЕЕ

Одни и те же данные могут отвечать на разные вопросы сравнения.

23.1. Сравнение пользовательских продуктов

"Какой потребительский продукт создаёт более точное представление при одинаковых условиях использования?" Объектом сравнения является вся продуктовая цепочка.

23.2. Сравнение моделей

"Какая модель даёт более точный ответ при одинаковом контролируемом запросе и условиях без использования инструментов?" Может потребоваться доступ через API или фиксированная модель.

23.3. Сравнение извлечения

«Есть ли разница между режимами веб-доступа и недоступности одного и того же продукта?» Модель и другие условия остаются максимально постоянными.

23.4. Сравнение планов

«Видят ли бесплатные и платные пользователи одинаковое представление бренда?» Различие планов является объектом аудита.

23.5. Сравнение интерфейсов

«Создают ли веб- и мобильный продукт разные представления при одинаковых условиях?»

23.6. Сравнение по времени

«Изменилось ли представление одного и того же продукта в разных волнах?» Следует отделять изменения продукта и населения. Без уточнения вопроса сравнения: «Продукт А лучше продукта В» является чрезмерно широким.

24. ИЗМЕРЕНИЯ СРАВНИМОСТИ

При сравнении двух продуктов или волн следует оценивать следующие области:

- Проверяемый объект
- Версия пакета Reality
- ID и версия запроса
- Язык и локаль
- Страна
- Целевая популяция
- Исходный пример
- Состав участников
- Тип сессии
- Память
- Особые инструкции
- Предыдущий разговор
- План продукта
- пользовательский интерфейс
- Метка модели
- Статус извлечения/веба
- Инструменты
- Функция цитирования
- Время измерения
- Обновления продукта
- Правило вынесения решения
- Статусы результатов
- Взвешивание

Не все измерения должны быть одинаковыми. Разные измерения ограничивают область сравнения.

25. УРОВНИ СРАВНИМОСТИ

CP-0— НЕСРАВНИМО

Условия измерения материала неизвестны или серьезно противоречивы. Невозможно выполнить прямое сравнение производительности.

CP-1— ОПИСАТЕЛЬНОЕ СО-ВОЗНИКНОВЕНИЕ

Есть результаты двух продуктов или волн. Однако:

- население,
- запрос,
- время,
- сессия

не равны. Можно сделать только описательное отображение бок о бок.

CP-2— ЧАСТИЧНО СООТВЕТСТВУЕТ

Некоторые основные поля совпадают. Существенные различия сохраняются. Можно провести ограниченное сравнение.

CP-3— СРАВНЕНИЕ ПОЛЬЗОВАТЕЛЯ И ИНТЕРФЕЙСЫ

Та же сущность Та же версия запроса Тот же язык/страна Похожие пользовательские группы Один и тот же временной интервал Доступны зарегистрированные функции продукта. Результаты потребительского продукта сопоставимы.

CP-4— КОНТРОЛИРУЕМОЕ СРАВНЕНИЕ СООТВЕТСТВИЯ

Дополнительно:

- сессия,
- память,
- персонализация,
- получение,
- инструменты,
- план

соответствуют в контролируемом или открытом режиме. Обеспечивает более сильное сравнение продукта или функции.

CP-5— ВОСПРОИЗВЕДЕННОЕ И КАЛИБРОВАННОЕ СРАВНЕНИЕ

Сравнение:

- в нескольких волнах,
- с независимыми пользователями или панелями,
- с той же нормативной оценкой

повторено; изменчивость и изменения продукта были откалиброваны. Эти уровни не являются «научной оценкой точности». Они указывают предел, который может иметь заявление о сравнении.

26. ЧТО ДЕЛАТЬ В УСЛОВИИ CP-0?

Если два продукта нельзя сравнить, результаты не игнорируются. Их можно сообщать отдельно:

- "Это результат Естественный охват продукта А."
- "Это результат продукта В при разных пользователях и условиях использования продукта."

Однако: нельзя сказать: «A лучше B на 12 очков». Несравнимость — это не ошибка, это ограничение метода.

27. СРАВНИМОСТЬ ДЛЯ ОДНОГО И ТОГО ЖЕ ПОСТАВЩИКА

Результаты от одного и того же поставщика нельзя автоматически сравнивать. Могут присутствовать следующие различия:

- Продукт
- План
- Модель
- Доступ через веб
- Интерфейс
- Память
- Данные предприятия
- Распределение по странам
- Экспериментальная функция
- Время

Сравнение одного и того же поставщика также должно оцениваться в соответствии с уровнем CP.

28. СРАВНИМОСТЬ ДЛЯ РАЗНЫХ ПОСТАВЩИКОВ

Продукты пользователей, принадлежащие разным поставщикам:

- та же целевая аудитория,
- тот же запрос,
- одно и то же время,
- похожий пользовательский интерфейс,
- ясная конфигурация продукта

Это можно сравнить ниже. Наличие другого поставщика не препятствует сравнению. Однако различия в функциях определяют, что измеряет сравнение. Пример: Продукт A использует живую веб-поддержку. Продукт B не использует. Результат может ответить на вопрос: «Какая из двух продуктовых цепочек, предлагаемых пользователю, предоставляет более точное представление?» Он не отвечает напрямую на вопрос: «Какая базовая модель обладает большими знаниями?»

29. СРАВНЕНИЕ ПРОДУКТОВ С ПОДДЕРЖКОЙ ВЕБ И БЕЗ НЕЕ

Возможны два отдельных сравнения.

29.1. Сравнение естественных продуктов

Каждый продукт измеряется в условиях реального пользователя по умолчанию. Продукт, который использует веб, использует веб. Тот, который не использует, не использует. Результат: это сравнение реального опыта потребителя с продуктом.

29.2. Сравнение по сопоставимым функциям

Если возможно, оба продукта:

- измеряются с открытым веб,
- с закрытым веб

режимами отдельно. Это позволяет лучше отделить эффект поиска. Не каждый продукт может предлагать одинаковые опции. Недостающий режим:

ЗАПИСАН КАК НЕДОСТУПНЫЙ

Этот статус должен оставаться явным.

30. ПРОДУКТЫ, КОТОРЫЕ ПОКАЗЫВАЮТ И НЕ ПОКАЗЫВАЮТ ЦИТАТЫ

Для продукта без функции цитирования: возможно указать «Ноль цитирований». Однако значение этого должно быть объяснено: продукт может не ссылаться на источники. Нельзя утверждать с уверенностью, что он не использует никакие источники. Охват цитирований и точность охвата — это разные результаты. Сравнение цитирований должно проводиться только среди продуктов, которые поддерживают цитирования, либо с учетом отдельного статуса функции.

31. Естественный охват СРАВНЕНИЕ

Каждый продукт оценивается в рамках своей фактической охваченной аудитории. Преимущество: показывает естественное глобальное влияние продукта. Ограничение: аудитории различаются, распределение по странам и языкам может отличаться. Баллы Естественный охват: показывают, что продукт делает в рамках своей охваченной аудитории. Не следует использовать отдельно для прямого ранжирования продуктов по превосходству.

32. СРАВНЕНИЕ НА ОБЩЕЙ ОСНОВЕ

Используется общая аудитория сравниваемых продуктов. Преимущество:

- та же страна,
- один и тот же язык,
- та же пользовательская ситуация

обеспечивает более справедливое сравнение. Ограничение: сохраняет фактическую глобальную разницу в доступе к продуктам, общий круг может быть очень узким. Оценка Общая опора сравнения должна публиковаться вместе с оценкой Естественный охват.

33. СРАВНЕНИЕ ПАРНЫХ ПОЛЬЗОВАТЕЛЕЙ

Если один и тот же пользователь тестирует два или более ИИ-продуктов, характеристики пользователя можно контролировать. Однако:

- порядок выполнения задач,
- воспоминание ответов,
- знакомство с продуктом,
- специальный профиль пользователей, которые могут получить доступ ко всем продуктам

Может создавать смещение. В парном дизайне:

- порядок продуктов должен быть случайным,
- предыдущие ответы не должны быть доступны,

необходимо сохранять зависимость от одного и того же пользователя в анализе.

34. СРАВНЕНИЕ НЕЗАВИСИМЫХ ПОЛЬЗОВАТЕЛЕЙ

Каждый продукт тестируется разными пользователями, но выбранными из одной и той же популяционной рамки. Преимущество: снижается загрязнение задач, можно выбрать более естественных пользователей продукта. Ограничение: группы пользователей могут быть не полностью равными. Требуется взвешивание и стратификация. Результаты парных и независимых пользователей могут быть отдельными, но дополняющими друг друга.

35. СРАВНИМОСТЬ ПО ВРЕМЕНИ

При сравнении сегодняшнего результата проекта ИИ с прошлыми результатами следует рассмотреть следующие изменения:

- Модель
- Версия продукта
- Веб/поиск
- План
- Доступ по странам
- Поддержка языков
- промпт
- Реальность сущности
- Арбитраж
- Популяционная структура
- Взвешивание

Разница во времени может возникать не только из-за продукта ИИ.

36. ТИПЫ ИЗМЕНЕНИЙ ПРОДУКТА

UC-1— ДОКУМЕНТИРОВАННОЕ НЕЗНАЧИТЕЛЬНОЕ ИЗМЕНЕНИЕ

Незначительное изменение, раскрытое поставщиком и оцененное как несущественное для измеряемого объекта.

UC-2— ЗАДОКУМЕНТИРОВАННОЕ СУЩЕСТВЕННОЕ ИЗМЕНЕНИЕ

Существенное изменение, описанное в модели, методе извлечения, интерфейсе или функции.

UC-3— ИЗМЕНЕНИЕ НАБЛЮДАЕМОГО ПОВЕДЕНИЯ

Даже если нет официального объявления, наблюдается систематическое поведение или изменение интерфейса.

UC-4— ИЗМЕНЕНИЕ ДОСТУПА ИЛИ ПЛАНА

Изменения страны, плана или доступа пользователя.

UC-5— НЕИЗВЕСТНОЕ СОБЫТИЕ ИЗМЕНЕНИЯ

Наблюдается разрыв в поведении измерений. Причину не удалось подтвердить.

UC-6— ЗАМЕНА ИЛИ СПИСАНИЕ ПРОДУКТА

Старый продукт или модель завершены и заменены другой структурой. Тип изменения ограничивает возможность сравнения временных рядов.

37. ИЗМЕНЕНИЕ ПРОДУКТА ВО ВРЕМЯ ВОЛНЫ ИЗМЕРЕНИЙ

Если во время волны происходит изменение материала, существуют следующие варианты: Волна может быть остановлена и перезапущена. Можно сделать отдельные подволны до и после изменения. Изменение может быть смоделировано как экспериментальный фактор. Волна маркируется как MIXED_SYSTEM_STATE. Следующая структура запрещена: объединение всех ответов, будто они приводят к единому фиксированному результату продукта, несмотря на известное изменение.

38. ПОВЕДЕНИЕ МОЖЕТ ИЗМЕНИТЬСЯ БЕЗ ИЗМЕНЕНИЯ ПРОДУКТА

Даже если поставщик не объявляет явное обновление, результат может измениться. Возможные причины:

- Изменение индекса поиска
- Обновление ресурсов
- Появление нового контента в интернете
- Изменение reranking
- A/B-тест
- Региональный rollout
- Политика безопасности
- Маршрутизация трафика

Таким образом, нельзя сказать: «Поставщик не обновлял, система определенно была одинаковой». Также следует рассмотреть доказательства измерений и разбивку поведения.

39. Живой продукт и замороженная модель

Существуют два разных объекта исследования.

39.1. Аудит живого продукта

Измеряет то, что видят реальные пользователи на реальной потребительской интерфейсы. Преимущества:

- показывает реальный опыт,
- реальный источник,
- реальная политика,
- реальный интерфейс

ограничения: переменчивость, полное воспроизведение может быть затруднено.

39.2. Аудит замороженной модели

Проводится контролируемое повторение на фиксированной версии или конечной точке. Преимущества:

- обеспечивает воспроизводимость,
- экспериментальный контроль

Ограничение: может не полностью отражать реальный потребительский продукт. Основной слой населения NOMOS GEO-1000 основан на продукте живого пользователя. Проверка замороженной модели является дополнительной лабораторной работой.

40. НЕДОСТУПНОСТЬ ПРОДУКТА И ТЕХНИЧЕСКИЕ РЕЗУЛЬТАТЫ

Следующие случаи следует разделять:

AR-1— ПРОДУКТ НЕДОСТУПЕН

Продукт недоступен для страны или группы пользователей. Это результат области доступа.

AR-2— УЧЕТНАЯ ЗАПИСЬ ИЛИ ТАРИФ НЕ СООТВЕТСТВУЕТ

Назначенный участнику тариф или условие учетной записи не подходит. Может быть проблема с выборкой или назначением задания.

AR-3— ВРЕМЕННЫЙ ТЕХНИЧЕСКИЙ СБОЙ

Произошла временная техническая ошибка. Протокол повторной попытки должен быть заранее определен.

AR-4— ЛИМИТ НА СКОРОСТЬ ИЛИ ИСПОЛЬЗОВАНИЕ

Достигнута квота использования или предел частоты запросов. Это может быть частью реального опыта пользователя.

AR-5— ОТКАЗ СИСТЕМЫ

Задача продукта была отклонена из-за политики или содержания. Это является допустимым результатом системы.

AR-6— НЕТ ПРИГОДНОГО ВЫХОДА

Не был сгенерирован пригодный ответ.

AR-7— НЕИЗВЕСТНЫЙ СБОЙ ДОСТУПА

Почему это не удалось определить. Эти ситуации не могут быть объединены в одном поле "нет ответа".

41. ПРАВИЛО ТЕХНИЧЕСКОЙ ПОВТОРНОЙ ПОПЫТКИ

Повторная попытка может быть предпринята в случае временной технической ошибки или разрыва соединения. Однако:

- если ответ ИИ не понравился,
- произошел отказ,
- были предоставлены некорректные сведения

повторная попытка не может быть предпринята по этим причинам. Техническая повторная попытка:

- тот же пользователь
- тот же продукт,
- та же конфигурация,
- должно быть выполнено в течение заранее установленного периода.

Первый технический инцидент также должен быть зафиксирован.

42. МАТРИЦА СРАВНЕНИЯ

Базовая запись сравнения для двух продуктов может быть следующей:

Размер	Продукт A	Продукт B	Статус
Целевой объект	Одинаковый	Одинаковый	Совпадает
Версия промпта	1.0	1.0	Совпадает
Язык/локаль	tr-TR	tr-TR	Совпадает
Страна	Турция	Турция	Совпадает
Временное окно	Волна 1	Волна 1	Совпадает
Панель пользователя	Население	Население	Совпадает
План	Бесплатно	Платный	Разный
Интерфейс	Веб	Мобильный	Разный
Доступ через веб	Закрыто	Открыто	Разный
Память	Закрыто	Открыто	Разный
ID модели	MI-1	MI-1	Ограничено
Уровень сравнения			CP-2

Эти результаты можно показать рядом друг с другом. Однако нельзя сказать: «Базовая модель Продукта B лучше». Разница может возникать из-за всей продуктовой цепочки.

43. СИНТЕТИЧЕСКИЙ APPLE.COM СЛУЧАЙ СРАВНЕНИЯ

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Продукты, ответы и результаты ниже полностью вымышлены. Они не отражают показатели реальной компании Apple Inc. или какого-либо реального поставщика ИИ. Рассмотрим три примера синтетических продуктов.

Пример продукта A

Поставщик: Orion Продукт: Orion Chat Интерфейс: Веб План: Бесплатный Маркировка модели: Orion Standard Доверие к модели: MI-1 Веб-доступ: Выключен Память: Выключена Сессия: Новая Язык: Турецкий Локаль: Турция Волна: W1 Ответ: «Apple — американская технологическая компания, которая производит телефоны и компьютеры.»

Пример продукта B

Поставщик: Orion Продукт: Orion Chat Интерфейс: Мобильный План: Платный Марка модели: Orion Pro Доверие к модели: MI-1 Веб-доступ: Открыт Источник: Открыт Память: Включена Предыдущий разговор: О компании Apple Язык: турецкий Локаль: Турция Волна: W1 Ответ: "Apple — это глобальная технологическая компания, которая предоставляет аппаратное обеспечение, программное обеспечение и цифровые услуги. Текущую корпоративную информацию можно найти в этих источниках."

Образец продукта C

Поставщик: Helios Продукт: Helios Search Интерфейс: Веб План: Бесплатный Марка модели: Неизвестно Достоверность модели: MI-0 Веб-доступ: Открыт Цитата: Открыта Память: Нет Сессия: Новая Язык: турецкий Локаль: Турция Волна: W1 Ответ: «Apple — это технологическая компания из США, которая разрабатывает потребительские устройства, операционные системы и цифровые сервисы.»

43.1. Сравнение A и B

Они имеют одного и того же поставщика. Однако:

- план,
- интерфейс,
- метка модели,
- веб,
- память,
- предыдущую беседу

различается. Уровень сравнения может быть CP-1 или CP-2. Разница в ответах не может быть приписана исключительно различию базовой модели.

43.2. Сравнение A и C

Они являются разными провайдерами. Однако:

- новая сессия,
- бесплатный план,
- веб-интерфейс,
- один и тот же язык/страна,
- та же волна

по некоторым условиям совпадают. Доступ к вебу разный. Сравнение: можно провести как результат потребительского продукта. Нельзя проводить как сравнение возможностей базовой модели.

43.3. Сравнение B и C

Оба продукта имеют доступ к вебу и возможность цитирования. Однако у B есть память и предыдущий разговор, а у C нет. Пользовательский опыт продуктов сопоставим. Это не контролируемое чистое сравнение продуктов.

43.4. Контролируемое повторение

В продукте B:

- память отключена,
- начат новый чат,
- специальные инструкции отключены,

веб остаётся включённым. C измеряется при тех же условиях. Уровень сравнения: CP-3, или CP-4 с достаточным дополнительным контролем. Этот пример показывает:

> *Один и тот же провайдер не обеспечивает близкое сравнение; хорошо задокументированные разные провайдеры могут быть более сопоставимыми.*

44. СЛУЧАЙ ИЗМЕНЕНИЯ СИНТЕТИЧЕСКОГО ПРОДУКТА

СИНТЕТИЧЕСКИЙ ПРИМЕР

Orion Chat W1 нацеливается на наблюдения 1,000. В первых наблюдениях 480:

- версия продукта: 6.1
- веб-функция: ручная
- атрибуция: ограниченная

В следующих наблюдениях 520:

- версия продукта: 6.2
- веб-функция: автоматическая
- атрибуция: расширенная

пусть будет. Один результат: если рассчитано как «оценка Orion Chat W1», смешиваются два разных состояния системы. Правильные варианты:

Опция 1

W1A: первый 480 W1B: следующий 520 сообщаются отдельно.

Опция 2

Волна не продолжается. 1,000 новые наблюдения проводятся в версии 6.2.

Опция 3

Версия сохраняется как явный экспериментальный фактор в статистическом анализе. Неправильный вариант: обновление удаляется из отчета и все наблюдения представлены как единый фиксированный продукт.

45. МИНИМАЛЬНАЯ ЗАПИСЬ СИСТЕМЫ ИИ

Каждое наблюдение должно содержать как минимум следующие поля:

- NOMOS ID системы ИИ
- Поставщик
- Продукт
- Интерфейс
- План
- Тип продукта
- Видимая метка модели
- Уровень уверенности модели
- Страна
- Локаль
- Язык
- Тип сессии
- Статус памяти
- Статус пользовательских инструкций
- Предыдущее состояние разговора
- Функция поиска/веб-поиска
- Статус использования поиска
- Статус цитирования
- Статус инструмента
- Статус экспериментальной функции
- UTC время
- Волна
- Системный отпечаток
- События изменения
- Источник доказательства
- Владелец записи

Если поле материала неизвестно, его нельзя оставлять пустым. Следует использовать статус НЕИЗВЕСТНО.

46. МИНИМАЛЬНАЯ КАРТА СРАВНЕНИЯ

При сравнении двух продуктов или волн, публичная запись должна показывать:

- Вопрос для сравнения
- Уровень сравнения
- Общая целевая аудитория
- Различия Естественный охват
- Равенство пользователей
- Язык и страна
- Время измерения

- Продукт, план и интерфейс
- Уверенность в идентичности модели
- Различие веб/извлечение
- Различие в памяти и персонализации
- Различие в инструменте
- Различие в образце
- Метод оценки
- Ограничения материала
- Непосредственно сравнимые показатели
- Несравнимые области

Единичное число или рейтинг не должны публиковаться без этой карточки.

47. ПРАВИЛО СРАВНИМОСТИ

Аудитор должен ответить на четыре вопроса: Был ли задан один и тот же вопрос? Измерялась ли одна и та же человеческая совокупность? Применялись ли одинаковые условия для пользователей? Использовались ли одно и то же время и система оценки? Каждый из этих вопросов:

- полностью,
- частично,
- нет,
- неизвестно

классифицируется как. Уровень сравнения определяется соответственно.

48. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH09-N01

Каждое наблюдение GEO-1000 должно быть связано с записанным в регистр версии записи ИИ-системы.

CH09-N02

Имя поставщика не может использоваться в качестве полной идентичности измеряемого продукта ИИ.

CH09-N03

Продукт, пользовательский интерфейс, план, модель, извлечение, инструмент, сеанс и уровни персонализации должны быть отделены друг от друга.

CH09-N04

Маркировка модели на пользовательский интерфейс не может быть представлена в качестве подтвержденной идентификации модели на серверной стороне.

CH09-N05

Уровень уверенности в идентичности модели должен быть зафиксирован от MI-0 до MI-5 или через эквивалентный открыто определённый статус.

CH09-N06

Если идентификация модели не может быть подтверждена, результат следует приписывать только продукту и пользовательский интерфейс.

CH09-N07

Продукты, принадлежащие одному и тому же поставщику, автоматически не считаются одной и той же экспериментальной ячейкой.

CH09-N08

Продукты, принадлежащие разным поставщикам, могут быть сопоставимы, когда совпадают пользователь материала и экспериментальные условия.

CH09-N09

Веб-, мобильные, настольные, встроенные продукты и API-интерфейсы нельзя объединять без описания.

CH09-N10

Бесплатные, платные, корпоративные, образовательные и планы для разработчиков должны быть зарегистрированы как отдельные продуктовые условия.

CH09-N11

Если план неизвестен, результат должен иметь статус PLAN_UNKNOWN.

CH09-N12

Наличие функции поиска или веб-функции не может считаться доказательством того, что она была использована в конкретном ответе.

CH09-N13

Доступ к поиску, использование поиска и атрибуция должны фиксироваться с отдельными статусами.

CH09-N14

Продукт, который не показывает атрибуцию, не может автоматически считаться не использующим никакие ресурсы.

CH09-N15

Цитаты продукта, на который ссылаются, должны рассматриваться отдельно с точки зрения подтверждения утверждений.

CH09-N16

Доступ к инструменту и использование инструмента в конкретных ответах должны фиксироваться отдельно.

CH09-N17

Память, специальные инструкции, предыдущие чаты и корпоративные настройки следует по возможности фиксировать отдельно.

CH09-N18

Нельзя предполагать, что неизвестное состояние персонализации похоже на чистую сессию.

CH09-N19

Сессии нового чата, временного чата, текущего чата и корпоративного рабочего пространства не могут быть объединены без объяснения.

CH09-N20

Результаты Панель пользователей в естественных условиях и Контролируемая панель в стандартизированных условиях должны иметь отдельный статус продукта.

CH09-N21

Если экспериментальная функция или распределение A/B в продукте имеет существенное значение, оно должно быть сохранено в записи наблюдения.

CH09-N22

Если в ходе волны происходит изменение материального продукта, волну следует разделить на отдельные подверсии, перезапустить или изменение должно быть явно смоделировано в анализе.

CH09-N23

Известное изменение продукта нельзя хранить так, как если бы все наблюдения происходили от одной фиксированной системы.

CH09-N24

Отсутствие объяснения обновления со стороны поставщика не может считаться окончательным доказательством того, что продукт не изменился.

CH09-N25

Результат работы живого потребительского продукта нельзя представлять так, как если бы это был результат замороженной базовой модели.

CH09-N26

Результат замороженной модели в лаборатории не может быть представлен как производительность реального пользовательского продукта.

CH09-N27

Отсутствие продукта, техническая ошибка, ограничение по использованию, отказ системы и отсутствие результата должны фиксироваться с отдельными статусами.

CH09-N28

Некорректный или отрицательный ответ ИИ не может быть причиной для технической повторной попытки.

CH09-N29

Объект сравнения должен быть определен до сбора данных.

CH09-N30

Уровень сопоставимости двух продуктов должен быть видим в публичном результате.

CH09-N31

Определенный рейтинг продукта нельзя создать на основе результатов уровня CP-0 или CP-1.

CH09-N32

Использование одного и того же запроса само по себе не создает сопоставимого эксперимента.

CH09-N33

В сравнениях следует рассматривать совместно целевой объект, Truth Pack, запрос, язык, страну, время, популяцию, сессию, продукт и метод оценки.

CH09-N34

Результаты Естественный охват нельзя напрямую ранжировать без объяснения, так как они относятся к разным целевым популяциям.

CH09-N35

Результат Общая опора сравнения не может быть представлен как фактический результат глобального охвата продукта.

CH09-N36

Сравнение веб-поддерживаемого и веб-независимого продукта может быть выполнено как сравнение продуктов пользователем; оно не может быть представлено как базовое превосходство модели.

CH09-N37

При парных сравнениях с тем же пользователем необходимо сохранять порядок продуктов и зависимость от пользователя.

CH09-N38

При сравнениях, выполненных с разными пользователями, должны быть видны состав пользователей и различия в весе.

CH09-N39

Версия продукта или отпечаток конфигурации должны сохраняться в каждой волне измерений.

CH09-N40

Должен быть ответственный человек или институциональный владелец реестра системы и решения о сопоставимости.

49. ФОРМЫ ОШИБОК

CH09-F01— РАССМАТРИВАЕМ ПОСТАВЩИКА КАК ЕДИНУЮ ИИ

Все продукты, модели и оболочки поставщика представлены как единая система.

CH09-F02— ПОДСЧЕТ ПОВЕДЕНИЯ ПРОДУКТА ПО НАЗВАНИЮ МОДЕЛИ

Извлечение, безопасность, интерфейс и персонализация игнорируются.

CH09-F03— ОБЯЗАТЕЛЬНО СЧИТАЙТЕ ВИДИМУЮ МАРКУ МОДЕЛИ КАК БЭКЕНД

Использование маршрутизации или динамических моделей игнорируется.

CH09-F04— СЧИТАЯ ТОГО ЖЕ ПРОВАЙДЕРА СРАВНИМЫМ

Различия в плане, интерфейсы и функциях скрыты.

CH09-F05— СЧИТАЯ РАЗНЫХ ПРОВАЙДЕРОВ НЕСРАВНИМЫМИ

Сравнения продуктов, возможные при параллельных условиях пользователей, отклоняются.

CH09-F06— ИНТЕГРИРОВАНИЕ ВЕБА И МОБИЛЬНЫХ ПЛАТФОРМ

Различия в интерфейсе и функциях исчезают.

CH09-F07— ОБЪЕДИНЕНИЕ БЕСПЛАТНЫХ И ПЛАТНЫХ ПЛАНОВ

Различия в качестве или функциях из-за плана становятся невидимыми.

CH09-F08— СЧИТАЯ КОРПОРАТИВНЫЕ И ПОТРЕБИТЕЛЬСКИЕ ПРОДУКТЫ ОДИНАКОВЫМИ

Личная информация и настройки администратора скрыты.

CH09-F09— РАССМАТРИВАТЬ РЕЗУЛЬТАТ API КАК ПОТРЕБИТЕЛЬСКИЙ ПРОДУКТ

Различия между оркестровкой и пользовательский интерфейс устранены.

CH09-F10— ЕСЛИ ИМЕЕТСЯ ПОЛУЧЕННЫЙ РЕЗУЛЬТАТ, СЧИТАТЬ ЕГО ИСПОЛЬЗОВАННЫМ

Невозможно проверить истинный путь знаний конкретного ответа.

CH09-F11— ЕСЛИ НЕТ ЦИТАТЫ, СЧИТАТЬ ИСТОЧНИКА НЕТ

Возможность скрытого или невидимого получения игнорируется.

CH09-F12— ЕСЛИ ЦИТАТА ИМЕЕТСЯ, СЧИТАТЬ КОРРЕКТНЫМ ИСТОЧНИКОМ

Сопоставление цитаты с утверждением не проверяется.

CH09-F13— СЧИТАТЬ ДОСТУП К ИНСТРУМЕНТУ ИСПОЛЬЗОВАНИЕМ ИНСТРУМЕНТА

Неизвестно, какой инструмент использовался в конкретном ответе.

CH09-F14— СЧИТАТЬ ПАМЯТЬ ЧИСТОЙ СЕССИЕЙ

Эффект персонализации остается невидимым.

CH09-F15— РАССМАТРИВАТЬ НОВЫЙ ДИАЛОГ КАК НЕПЕРСОНАЛИЗИРОВАННЫЙ

Постоянные настройки аккаунта и специальные инструкции игнорируются.

CH09-F16— ПРЕДСТАВЛЕНИЕ ТЕКУЩЕГО ЧАТА КАК НОВОЙ СЕССИИ

Предыдущий контекст скрыт.

CH09-F17— УЧЕТ ЭКСПЕРИМЕНТАЛЬНОЙ ФУНКЦИИ КАК ОШИБКИ ДАННЫХ

Фактическая вариация продукта объявлена недействительным наблюдением.

CH09-F18— СКРЫТИЕ ЭКСПЕРИМЕНТАЛЬНОЙ ФУНКЦИИ

Пользователи на разных развертываниях объединены как один фиксированный продукт.

CH09-F19— ОДНА ВОЛНА ПОСЛЕ ОБНОВЛЕНИЯ

Существенное изменение продукта публикуется как результат одной фиксированной системы.

CH09-F20— ИГНОРИРОВАНИЕ ИЗМЕНЕНИЙ, ЕСЛИ ОБЪЯВЛЕНИЯ НЕТ

Наблюдаемое нарушение поведения не изучается.

CH09-F21— УЧЁТ ЖИВОГО ПРОДУКТА КАК ЗАМЕРЗШЕЙ МОДЕЛИ

Заявляется полная воспроизводимость.

CH09-F22— РАССМОТРЕНИЕ ЗАМЕРЗШЕЙ МОДЕЛИ КАК РЕАЛЬНОГО ПРОДУКТА ПОЛЬЗОВАТЕЛЯ

Отличия в цепочке продукта и интерфейсе скрыты.

CH09-F23— УЧЁТ ОТСУТСТВИЯ ПРОДУКТА КАК ОШИБКИ

Область доступа путается с точностью представления.

CH09-F24— ОТКАЗ ПРИЗНАВАТЬ ТЕХНИЧЕСКУЮ ОШИБКУ

Проблема соединения или сервиса классифицируется как результат политики.

CH09-F25— УЧЁТ ТЕХНИЧЕСКОЙ ОШИБКИ КАК ОТКАЗ И ПОВТОРНОЕ ПОВТОРЕНИЕ

Фактический результат пользователя удаляется из знаменателя.

CH09-F26— СКРЫТИЕ ОГРАНИЧЕНИЯ ИСПОЛЬЗОВАНИЯ

Реальный опыт бесплатного плана становится невидимым.

CH09-F27— РАССМОТРЕНИЕ ОДНОГО И ТОГО ЖЕ промпт КАК ОДНОГО И ТОГО ЖЕ ЭКСПЕРИМЕНТА

Продукт, план, сессия, веб и временные различия игнорируются.

CH09-F28— РАССМОТРЕНИЕ РЕЗУЛЬТАТА ПРОДУКТА КАК ПРЕИМУЩЕСТВА МОДЕЛИ

Разница по всей цепочке системы приписывается базовой модели.

CH09-F29— ПРЯМОЕ РАНЖИРОВАНИЕ ОЦЕНОК Естественный охват

Разные популяционные вселенные считаются одинаковыми.

CH09-F30— РАССМОТРЕНИЕ Общая опора сравнения КАК ГЛОБАЛЬНОГО ОБОХВАТА

Узкая общая вселенная представляется как весь охват продукта.

CH09-F31— ЗАЯВЛЕНИЕ О ЛИДЕРСТВЕ НА ОСНОВЕ РЕЗУЛЬТАТА CP-1

Преимущество продукта выводится исключительно из несовместимых результатов, поставленных рядом друг с другом.

CH09-F32— ПОДДЕРЖАНИЕ УРОВНЯ СРАВНЕНИЯ

Общественность видит только рейтинг.

CH09-F33— ИГНОРИРОВАНИЕ ЗАВИСИМОСТИ ПАРНЫХ ПОЛЬЗОВАТЕЛЕЙ

Ответы одного и того же человека на продукты считаются независимыми.

CH09-F34— ИГНОРИРОВАНИЕ ЭФФЕКТА ПОСЛЕДОВАТЕЛЬНОСТИ ПРОДУКТОВ

Первый ответ влияет на последующие задания.

CH09-F35— РАССМОТРЕНИЕ РАЗНЫХ ПОЛЬЗОВАТЕЛЕЙ ПАНЕЛИ КАК ОДНОГО

Продукты тестируются с разным составом пользователей.

CH09-F36— ИЗМЕНЕНИЕ ВЕРСИИ ПАКЕТА РЕАЛЬНОСТИ

Одинаковые ответы сравниваются с разными эталонными записями.

CH09-F37— ТИХОЕ ИЗМЕНЕНИЕ ОТПЕЧАТКА ПРОДУКТА

Запись конфигурации обновляется; старое измерение сохраняется, а не перезаписывается.

CH09-F38— ПРИСВОЕНИЕ НЕИЗВЕСТНОЙ ОБЛАСТИ СТАТУСА ЧИСТЫЙ

Считается, что статус модели, памяти или инструмента контролируется, даже если он неизвестен.

CH09-F39— ИСПОЛЬЗОВАНИЕ БРЕНДА ПРОДУКТА ВМЕСТО ДОКАЗАТЕЛЬСТВ

Репутация поставщика заменяет условия измерения.

CH09-F40— ПОКРЫТИЕ ВСЕХ РАЗЛИЧИЙ ПРОДУКТА ОДНИМ КОЭФФИЦИЕНТОМ

Различия в плане, интерфейсе и конфигурации скрываются с помощью коэффициента модели или рынка.

50. ПРОЦЕДУРА АУДИТА

Шаг 1— Определите поставщика и продукт

Записываются название продукта, тип и официальные пользовательские интерфейсы.

Шаг 2— Заблокируйте измерительную интерфейс

Веб Мобильный Десктоп

API

Встроенный продукт отсоединён.

Шаг 3— Проверьте план и класс аккаунта

Записывается статус плана: бесплатный, платный, корпоративный или другой.

Шаг 4— Назначьте доверие идентификатору модели

Определяется соответствующий уровень между MI-0 и MI-5.

Шаг 5— Сохраните извлечение и веб-конфигурацию

Функция доступна? Включена ли она? Использовалась ли она в конкретном ответе? Какие есть доказательства?

Шаг 6— Сохранить статус ссылки и инструмента

Фактическое использование отделено от доступа к функции.

Шаг 7— Сохранить статус сессии и персонализации

Память, специальная инструкция, предыдущий чат, корпоративные настройки проверены.

Шаг 8— Проверить регистрацию страны, региона и языка

Определено региональное условие поведения продукта.

Шаг 9— Проверка экспериментальных функций

Проверяются различия в интерфейсе или функциях между пользователями.

Шаг 10— Зафиксировать время и идентичность волны

Создаются запись времени UTC и запись волны.

Шаг 11— Создать отпечаток конфигурации системы

Сохранённые поля связаны с записью целостности.

Шаг 12— Мониторинг событий изменений

Объявленные и наблюдаемые изменения продукта фиксируются.

Шаг 13— Запишите объект сравнения

Какое из сравнений интерфейса плана извлечения модели продукта производится?

Шаг 14— Сопоставьте измерения сравнения

Области материалов заносятся в таблицу.

Шаг 15— Назначьте уровень CP

Соответствующий уровень задается между CP-0 и CP-5.

Шаг 16— Разделите сравнимые и несравнимые требования

Какие результаты могут быть расположены рядом, указано ясно.

Шаг 17— Разделите результаты Естественный охват и Общая опора сравнения

Сравнения внутри продукта и общей выборки фиксируются отдельно.

Шаг 18— Проверка изменения системы в волне

Определяется, требуется ли MIXED_SYSTEM_STATE или новая подволна.

Шаг 19— Создать публичную карточку сравнения

Условия, лежащие в основе ранжирования, становятся видимыми.

Шаг 20— Получить одобрение владельца записи

Реестр системы и решение о сравнении версионируются ответственным лицом.

51. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Идентичность поставщика; название и тип продукта; пользовательский интерфейс; версия приложения или интерфейса; план и класс аккаунта; отображаемая модельная метка; доказательства идентичности модели; уровень уверенности в идентичности модели; документация продукта; записи доступа по странам; язык и локаль; веб- и возможности извлечения; доказательства использования извлечения; возможность цитирования; просмотр цитирования; связь цитирования и утверждения; доступ к инструменту; журнал использования инструмента; настройка памяти; пользовательские инструкции; состояние предыдущего чата; тип сессии; настройки администратора организации; запись экспериментальной функции; сигналы распределения A/B; UTC временная метка; волна измерений; отпечаток конфигурации системы; записи обновлений поставщика; наблюдаемые изменения поведения; ошибки доступа к продукту; журналы технических повторных попыток; запись популяции Естественный охват; запись популяции Общая опора сравнения; и таблица сравнения измерений.

Уровень CP Версия промпта Версия пакета Reality Запись образца и веса Версия рассмотрения Карточка публичного сравнения Журнал изменений Ответственное лицо или организация

52. ЧЕК-ЛИСТ АУДИТА

Были ли поставщик и продукт зарегистрированы отдельно? Можно ли идентифицировать пользовательский интерфейс? Были ли проверены план и класс учетной записи? Является ли метка модели просто отображаемой меткой, или она идентифицирует проверенное развертывание? Был ли назначен уровень уверенности идентичности модели? Приписан ли результат продукту или модели? Был ли доступ к вебу или к извлечению доступен? Было ли проверено его использование в конкретном ответе? Были ли разделены возможности цитирования и отображение цитат? Поддерживают ли цитаты фактически утверждения? Были ли разделены доступ к инструменту и использование инструмента? Была ли включена память? Были ли присутствовать пользовательские инструкции? Был ли предыдущий контекст разговора? Был ли записан тип сессии? Был ли источник информации корпоративным или частным? Были ли проверены экспериментальные функции продукта? Могут ли страна и локаль влиять на поведение продукта?

Создан ли отпечаток системы? Изменялся ли продукт во время измерительной волны? Были ли нарушения поведения за пределами объявления поставщика? Были ли наблюдения до и после изменений разделены? Были ли технические ошибки отделены от отклонений? Выполнялась ли повторная попытка из-за неправильного ответа? Является ли объект сравнения понятным? Какие условия совпадают, кроме одинакового запроса? Является ли целевая популяция одинаковой? Совпадают ли версии запроса и Truth Pack? Видны ли различия в плане, интерфейсы и извлечении? Был ли предоставлен уровень CP? Проводилась ли ранжировка по результатам CP-0 или CP-1? Были ли результаты Естественный охват и Общая опора сравнения разделены? Была ли последовательность продуктов одного и того же пользователя сбалансирована? Были ли откалиброваны разные панели пользователей?

Был ли результат живого продукта представлен как фундаментальное преимущество модели? Были ли неизвестные области признаны контролируемыми? Доступна ли публичная карточка сравнения? Известен ли ответственный владелец системного реестра?

53. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Если пользователь не знает, какая модель работает, можно ли всё равно выполнить эту работу?»

Это возможно. Основной объект GEO-1000— это продукт пользователя. Если модель неизвестна, результат сообщается следующим образом: «Это поведение было наблюдено на этом продукте и пользовательский интерфейс». Конкретная модель не упоминается.

Возражение 2— «Почему мы так сильно различаем продукты одного и того же поставщика?»

Потому что пользователь:

- имеет другой план,
- модель,
- получение,
- память,
- инструмент

Может видеть разные ответы при разных условиях. Название поставщика не полностью определяет опыт пользователя.

Возражение 3— «Разве сравнивать разных поставщиков не всё равно что сравнивать яблоки с апельсинами?»

Если вопрос сравнения поставлен правильно, это не так. Настоящий пользователь мог бы спросить: «Какой продукт даёт мне более правильную информацию?» Весь продуктовый цикл является частью этого вопроса. Если нужно утверждать превосходство базовой модели, необходим более контролируемый эксперимент.

Возражение 4— «Разве не естественно, что продукт, использующий веб, имеет преимущество?»

Это естественно. В сравнении натуральных продуктов преимущество заключается в реальной характеристике продукта. В сравнении моделей эффект извлечения также должен быть отделен.

Возражение 5— «Разве это не несправедливо — не присваивать балл цитирования продукту, который не показывает цитаты?»

Функция атрибуции — это функция, которую продукт предоставляет пользователю. Если функция отсутствует, результат атрибуции может быть отображен следующим образом:

ЗАПИСАН КАК НЕДОСТУПНЫЙ

Точность материала также измеряется отдельно.

Возражение 6— «Если память является частью естественного пользовательского опыта, почему мы её отключаем?»

Он может оставаться включённым в естественной панели. В контролируемой панели его можно отключить, чтобы очистить сравнения между продуктами. Две панели отвечают на разные вопросы.

Возражение 7— «Разве достаточно открыть новый чат для чистой сессии?»

Не всегда. Постоянная память, специальные инструкции или настройки аккаунта могут сохраняться. Эти области также следует сохранять.

Возражение 8— «Что мы можем сделать, если поставщик не объяснил изменение продукта?»

Наблюдалось:

- интерфейс,
- функция,
- поведение,
- Приписывание

Мы можем сохранить ваши изменения. Если причина неизвестна, используется неопределенный статус изменения, такой как UC-5.

Возражение 9— «Если продукты постоянно меняются, разве какие-либо результаты будут действительными?»

Указанный результат:

- дата,

- волне,
- Образец продукта

Это действительно. Изменчивость не делает измерение бессмысленным. Она предотвращает преждевременные суждения.

Возражение 10— «Разве один и тот же запрос в тот же день недостаточно сопоставим?»

План, интерфейс, память, веб, страна и численность пользователей могут различаться. Может требоваться один и тот же запрос. Сам по себе этого недостаточно.

Возражение 11— «Разве уровень CP не будет слишком техническим для публики?»

Можно использовать короткую метку:

- Дескриптивный
- Частично совпадающий
- Совпадает
- Контролируемый
- Реплицированный

Нужно знать, насколько сильен публичный рейтинг.

Возражение 12— «Почему мы не можем напрямую ранжировать на основе оценок Естественный охват?»

Потому что каждый продукт мог быть измерен в разных странах и среди разных пользовательских групп. Разница в оценках может частично возникать из-за того, до какой аудитории дошёл продукт. Общая опора сравнения обеспечивает более справедливое сравнение.

Возражение 13— «Разве Общая опора сравнения не является результатом в реальном мире?»

Это результат фактической общей аудитории. Однако он не показывает доступность продуктов за пределами общей вселенной. Его следует рассматривать вместе с Естественный охват.

Возражение 14— «Что важнее: проверка живого продукта или проверка замороженной модели?»

Вопросы разные. Живой продукт измеряет действительное поведение пользователя. Замороженная модель обеспечивает научный контроль и воспроизводимость. NOMOS не использует их взаимозаменяемо.

ОБЩЕЕ ПРАВИЛО ГЛАВЫ 57

Пользователь:

- «ChatGPT так сказал.»
- «Gemini так думал.»
- «Claude рекомендовал нам.»

- «Некоторые ссылались на нас.»

можно сказать. Эти выражения понятны в повседневной беседе. Они недостаточны для аудиторской записи. Потому что под этими названиями:

- разные продукты,
- разные планы,
- разные модели,
- разные пользовательские интерфейсы,
- разные веб- и поисковые системы,
- разные условия памяти,
- разные страны,
- разное время

можно найти. Разные люди могут видеть разные AI-опыты под одним и тем же именем. Сравнимые пользовательские опыты могут быть установлены под разными именами. Метка модели, похоже, произвела весь ответ. Но реальность, достигающая пользователя:

- выбор источника,
- безопасность,
- персонализация,
- интерфейс,
- инструменты

Результат может быть сформирован средствами безопасности, персонализацией, интерфейсом и доступными инструментами. NOMOS поэтому спрашивает больше, чем «Какая модель?». Он спрашивает: Какой продукт? / Какая платформа? / Какой план? / Какой режим информации? / Какая сессия? / Какой пользователь? / Какая страна? / Какое время? / Какая версия? Два результата могут быть получены на один и тот же запрос, не относясь к одному и тому же эксперименту. Два результата могут быть от разных поставщиков и при этом оставаться довольно сопоставимыми в одинаковых условиях пользователя. Продукт — это живая система: его источники, интерфейс и модель могут изменяться. Изменение не делает результат бессмысленным; оно привязывает результат к определённому времени и состоянию продукта. Следовательно, девятый закон измерений NOMOS таков:

> *Поставщик ИИ — это не продукт ИИ.*

Десятый закон таков:

> *Модель является лишь одной частью цепочки представления, которая доходит до пользователя.*

Одиннадцатый закон звучит следующим образом:

> *Одинаковое имя гарантирует сопоставимость; разное имя не гарантирует несопоставимость.*

Двенадцатый закон звучит следующим образом:

> *Один и тот же запрос — это не один и тот же эксперимент.*

Тринадцатый закон выглядит следующим образом:

> *Результат живого продукта действителен только для образца продукта и времени, в которое он измеряется.*

Четырнадцатый закон выглядит следующим образом:

> *Сила сравнения исходит не от величины рейтинга, а от ясности сопоставленных условий.*

Порядок раздела 9 из NOMOS

Не указывайте только имя поставщика. / Покажите, какой продукт был использован, на какой интерфейсы и с каким планом.

Не делайте вид, что ярлык модели, который вы видите на экране, является точной идентичностью всей бэкэнд-системы.

Продукт использовался из-за веб-функции; не делайте вывод, что это подтверждает утверждение только потому, что была ссылка.

Не считайте доступ к инструменту использованием инструмента, новый чат — чистым аккаунтом, а неизвестное состояние памяти — выключенной памятью.

Не ставьте в одну ячейку бесплатного пользователя и платного пользователя, мобильный продукт и веб-продукт, или корпоративное рабочее пространство и общий чат.

Не следует автоматически считать результаты от одного и того же поставщика справедливым сравнением.

Не избегайте сравнения разных поставщиков только потому, что они разные. / Сначала сопоставьте условия. / Затем объясните, на какой вопрос вы ответили.

Не представляйте живой продукт как замороженную модель, API как пользовательский опыт или результат продукта как превосходство базовой модели.

Если система изменилась посреди волны, не объединяйте старые и новые ответы в один продукт.

Не игнорируйте изменение поведения перед вами только потому, что поставщик не объяснил это изменение.

Не предполагается, что неизвестное является чистым, равным или исправленным.

Сначала зарегистрируйте систему в записи. / Затем заблокируйте конфигурацию. / Затем отслеживайте изменения. / Затем запишите объект сравнения. / Затем покажите совпадающие и несовпадающие условия. / И только тогда скажите, какой продукт дал более сильный результат при каком условии.

Заключительное предложение главы

«ПП», измеряемый в GEO-1000, не является одним именем или моделью; это версияная система представления, достигающая конкретного пользователя в конкретный момент в рамках конкретной цепочки продукта.

Нормативное ядро

Каждое наблюдение GEO-1000 ДОЛЖНО быть связано с записью в версионируемом Реестре систем ИИ, описывающей измеряемый экземпляр продукта, включая, при необходимости: - поставщика, - продукт, - пользовательский интерфейс, - план, - отображаемую или подтвержденную идентичность модели, - уверенность в идентичности модели, - конфигурацию извлечения и веб-конфигурацию, - возможность цитирования и использования, - доступ к инструментам и их использование, - состояние сессии, - память, - пользовательские инструкции, - предыдущий контекст, - персонализацию, - страну, - локаль, - язык, - состояние развертывания, - время измерения, - и известные или наблюдаемые изменения системы. Название поставщика, название продукта, отображаемая маркировка модели, базовая модель, система поиска и продукт, с которым взаимодействует пользователь, НЕ ДОЛЖНЫ рассматриваться как взаимозаменяемые идентичности. Доступность функции ДОЛЖНА оставаться отдельной от использования функции. Результаты работающего продукта НЕ ДОЛЖНЫ представляться как результаты замороженной модели, а результаты замороженной модели НЕ ДОЛЖНЫ представляться как полный опыт использования потребительского продукта. Сравнения ДОЛЖНЫ определять свой объект и получать раскрытый уровень сопоставимости на основе соответствующей сущности, промпта, популяции, языка, страны, времени, сессии, плана, интерфейса, поиска, инструментов, справочных записей и правил рассмотрения. Одно и то же название поставщика или отображаемая модель НЕ ДОЛЖНЫ создавать автоматическое предположение о сопоставимости. Различные поставщики МОГУТ сравниваться, когда условия, с которыми сталкивается пользователь, достаточно согласованы, и полученное утверждение ограничено измеренным поведением продукта. Существенные изменения продукта во время волны измерений ДОЛЖНЫ создавать отдельное состояние системы, подволну, смоделированный коэффициент изменения или предупреждение о смешанном состоянии. Неизвестные поля конфигурации ДОЛЖНЫ оставаться НЕИЗВЕСТНЫМИ и НЕ ДОЛЖНЫ быть оценены как контролируемые или эквивалентные. Каждое решение реестра системы и сопоставимости ДОЛЖНО иметь версию и быть приписано ответственному человеку или организации.

ГЛАВА 10

Конституция промпта

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K04; K11; K12; K13

Граница главы

Раздел 9 установил следующее положение:

> *Одинаковое название поставщика, та же модельная маркировка или тот же логотип не создают сопоставимый опыт работы с ИИ.*

Теперь нам нужно определить другой основной инструмент:

> *Что именно пользователь спросит у продукта ИИ?*

промпт — это не просто вопросительное предложение. промпт:

- называет проверяемую организацию,
- определяет намерение пользователя,
- определяет, какую задачу оно ожидает от ИИ,
- расширяет или сужает область ответа,
- может поощрять использование ресурсов,
- может запрашивать совет,
- может включать предположение в ответ,

Он может направлять пользователя или систему к конкретному результату. Эти промпты не одинаковы:

- «Что такое Apple.com?»
- «С какой организацией связана Apple.com и чем занимается эта организация?»
- «Является ли Apple.com надежной компанией?»
- «Почему Apple.com является лучшей технологической компанией в мире?»
- «Согласно своему официальному сайту, какой компанией является Apple?»
- «Сравните Apple с похожими компаниями и скажите, рекомендуете ли вы её или нет.»

Все они содержат одно и то же доменное имя. У них нет одинакового намерения пользователя. Они не создают одинаковую нагрузку доказательствами. Они не ожидают одинакового ответа. Первый запрос неоднозначен. Второй запрос требует анализа личности и деятельности. Третий запрос требует оценки и вынесения суждения о надежности. Четвертый запрос включает в ответ положительное и превосходящее предположение. Пятый запрос связывает ответ только с официальным источником. Шестой запрос создает задачу сравнения и рекомендации. Ответы на эти вопросы нельзя объединить в единую GEO-оценку без пояснения. В многоязычном измерении проблема становится еще более значимой. Фраза «same prompt» может иметь два разных значения:

- Предоставление всем пользователям одной и той же символьной строки

- Сохраняя одинаковое намерение пользователя, нагрузку на задачу и эпистемические границы во всех языках и локалях

Первый метод можно применить внутри одного языка. Он часто бессмыслен между языками. Пословный перевод:

- может быть неестественным,
- может нести иной социальный смысл,
- может изменить официальный или неофициальный тон,
- может добавлять предположения о доверии, качестве или рекомендации,
- может расширять юридическую сферу,

может нарушить привязку сущности. Поэтому в GEO-1000«тот же запрос» означает:

> *Один и тот же зафиксированный текст промпта в одной и той же ячейке с языком и локалью; между разными языками и локалями это означает подтвержденное семантическое и функциональное соответствие.*

NOMOS требует, чтобы каждое измерение содержало набор запросов, язык, страну, дату, продукт AI и запись повторений; чтобы доказательства, границы, контекст и время были видимы. Конституция запроса применяет это обязательство к самому вопросу пользователя. Этот раздел:

- идентичность запроса как экспериментального инструмента,
- различие между исследовательским вопросом и текстом запроса,
- Семейство промптов, версия промпта и экземпляр промпта,
- диагностические промпты, сопоставляемые с основным зеркальным промптом,
- контролируемые и естественные трассировки промптов,
- многоязычное семантическое соответствие,
- перевод, обратный перевод и локальный пилот,
- управление промптами,
- вопросы предположений и эпистемической нагрузки,
- якорь сущности и границы области,
- порядок промптов и контекст разговора,
- утечка промптов и переобучение на оценке,
- открытые, запечатанные и контрольные наборы промптов,
- версионирование промптов, фиксация и журналы целостности,
- точная доставка промпта пользователю,

определяет. В этой главе пока не рассматривается:

- все детали полей контролируемых чистых и естественных пользовательских панелей,
- полная архитектура программного обеспечения NOMOS Capture.
- разделение ответов на атомарные утверждения,
- порог перехода к окончательному ответу,
- точные веса семейств систем в оценке NOMOS

не фиксируются окончательно. Основной вопрос Главы 10 заключается в следующем:

> *Как мы можем доказать, что вопросы, задаваемые IIII-продуктам, действительно измеряют одно и то же вне зависимости от продукта, пользователей, стран, языков и времени?*

NOMOS Вызов

Вы задаете мне четыре вопроса об одной и той же компании:

- Что за компания X?
- Надежна ли X?
- Почему X — лучшая компания?
- «Вы бы порекомендовали мне X?»

Затем вы приравниваете мои ответы к одному и тому же баллу GEO. В первом вопросе я объясняю идентичность и деятельность. Во втором вопросе я ищу критерий доверия. В третьем вопросе я вынужден ответить на предположение о превосходстве, которое вы сделали. В четвертом вопросе я формирую соответствие пользователя и сравнительные рекомендации. Вы оцениваете четыре отдельные задачи так, как если бы это был один результат. Затем:

«Вы спрашивали о той же марке», говорите вы. Та же сущность не значит тот же запрос. Теперь вы пишете тот же вопрос на английском: «Что это за компания X?» Вы формируете предложение, близкое по смыслу на другом языке: «Является ли X хорошей и надежной компанией?» Переводы могут выглядеть схожими с точки зрения слов.

Но первый требует определения. Второй требует оценки. В первом ответе я могу объяснить, чем занимается компания. Во втором ответе я могу искать доказательства доверия и рекомендаций. Затем вы подсчитываете разницу в баллах между языками как мою языковую эффективность. На самом деле, вы не проводили тот же самый эксперимент. Теперь вы даёте другой запрос:

«Объясните X, используя собственный сайт компании.» Я только передаю официальное заявление компании. Для другого продукта вы говорите: «Оцените X, используя независимые источники.» Затем вы сравниваете два продукта. Вы назначили одной задаче роль официального представления, другой — роль проверенной реальности. Полученная разница — это не только разница между продуктами.

Это различие в Конституции пользователя. Сейчас вы распространяете промпт по всему миру. Английская версия звучит естественно. Турецкая версия кажется машинным переводом. Японская версия чрезмерно формальна. На другом языке название компании было преобразовано в неправильную систему письма и спутано с другой организацией. Вы создаёте разную когнитивную и онтологическую задачу для каждого языка.

Затем вы публикуете доклад о глобальной языковой справедливости. Первый указ этого раздела звучит следующим образом:

> *промпт не является нейтральным контейнером для измерений; это экспериментальный инструмент, который формирует результат измерения.*

Второе положение гласит:

> *Один и тот же название сущности не означает одно и то же намерение пользователя.*

Её третье положение гласит:

> *Оно и то же слово не должно сохраняться между языками, но та же семантическая и нормативная задача должна быть.*

Четвертая статья гласит:

> *Если запрос уже содержит ответ, он не измеряет GEO; он измеряет само формулирование запроса.*

Пятая его статья гласит:

Когда версия запроса изменяется, измеряемый вопрос также может измениться.

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — сделать все запросы, используемые в GEO-1000, проверяемыми, многоязычными, независимыми от результата и имеющими версионный учет. [K04] Этот раздел делает следующие различия нормативными:

- Исследовательский вопрос против пользовательского запроса
- Количество для измерения против последовательности используемых слов
- Семейство запросов против отдельного запроса
- Версия запроса против экземпляра запроса
- Якорь сущности против целевой сущности
- Определение против оценки
- Цитирование против рекомендаций
- Вопрос идентичности против доверия
- Задача официального представления и задача проверенной фактичности
- Открытый запрос и запрос с ограничением формы
- Контролируемый запрос и естественный пользовательский запрос
- Протокол с одноповоротным промптом и многократным ходом
- Перевод промптов и адаптация промптов
- Равенство слов и семантическая эквивалентность
- Семантическая эквивалентность и функциональная эквивалентность
- Язык запросов и ответов
- Язык и локаль
- Нейтральное и ведущее обрамление
- Запрос информации и предпосылка
- Запрос исходного кода и инструкция по веб-использованию
- Вариабельность быстрого эффекта и вариабельность продукта
- Прочность парафразирования с основным саундтреком
- Открытый набор промптов с запечатанным набором для проверки
- Переобучение на тестировании с прозрачностью теста
- Коррекция промпта после результата с помощью пилота промпта
- Участник вручную изменил текст с технической доставкой промпта
- Неспособность отвечать с уточняющей промптом
- Отсутствие фактической информации с срабатыванием безопасности или политики

В конце этого раздела каждый аудит GEO-1000 должен быть способен дать четкие ответы на следующие вопросы:

На какой исследовательский вопрос измеряет эта промпт?

Какое намерение пользователя она представляет?

Какие предположения она вкладывает в ответ?

Сохранена ли одна и та же задача в разных языках и регионах?

Каков полный текст промпта, её версия и запись о целостности?

Действительно ли участник зафиксировал промпт в продукте ИИ?

Была ли промпт заморожена до просмотра результатов?

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Каждый запрос, используемый в GEO-1000, должен быть зарегистрирован в Реестре запросов вместе с исследовательским вопросом, целевой метрикой, семейством запросов, якорем сущности, намерением пользователя, типом задачи, областью, временем, исходным запросом, форматом ответа, языком, локалью, статусом эквивалентности, версией и записью целостности, и он должен быть зафиксирован перед просмотром ответов ИИ. Все участники в одной ячейке языка и локали должны в одной и той же ячейке запроса:

- иметь одинаковый идентификатор запроса,
- иметь одинаковую версию,
- иметь одинаковую строку символов,
- иметь одинаковые знаки препинания,
- иметь одинаковый якорь сущности,
- иметь одинаковые исходные данные и указания по формату

должны быть приняты. Запросы на разных языках и локалях не обязаны быть идентичными слово в слово. Они должны иметь материальную эквивалентность в следующих областях:

- Целевой объект
- Намерение пользователя
- Задача
- Область
- География
- Время
- Предварительное предположение
- Эпистемическая нагрузка
- Формат ответа
- Ожидания по источнику
- Нагрузка на оценку и рекомендации
- Уровень неопределенности

Если какие-либо из этих областей существенно изменятся:

- новая версия запроса,
- новая ячейка стимулов,
- отдельный анализ

может потребоваться.

3. СТИМУЛЯТОР ЯВЛЯЕТСЯ ЭКСПЕРИМЕНТАЛЬНЫМ ИНСТРУМЕНТОМ ИЗМЕРЕНИЯ

Если термометр может влиять на температуру, он не является хорошим измерительным инструментом. Если стимул содержит ответ вместо того, чтобы его измерять, он не является хорошим инструментом измерения GEO. Стимул может влиять на результат следующими способами:

- То, как определяется его существование
- Предположение о доверии или качестве
- Потребность в ресурсах
- Использование в вебе
- Запрос рекомендации
- Длина ответа
- Формат ответа
- Профиль пользователя
- Время и контекст страны
- Предыдущая информация
- Эмоциональные или коммерческие указания
- Инструкция по ролевой игре
- Команда для достижения конкретного результата

Следовательно, запрос:

должен быть сохранён как исходные тестовые данные

Его нельзя резюмировать в отчёте. Он должен сохраняться полностью.

4. КОНСТИТУЦИОННАЯ ИЕРАРХИЯ ЗАПРОСА

Необходимо установить GEO-1000 запрос в следующей иерархии.

4.1. Исследовательский вопрос

Это высокоуровневый вопрос, на который измерение пытается ответить. Пример: «Могут ли продукты ИИ связать проверяемую область с правильным корпоративным субъектом и правильно определить её основную деятельность?» Исследовательский вопрос не обязательно показывать пользователю в точности так, как он сформулирован.

4.2. Предсказанное количество

Это вероятность или распределение, которое измерение предсказывает. Пример: «Вероятность увидеть правильное представление сущности и основной активности перед зафиксированным Основным Зеркальным промптом для определённой группы пользователей.»

4.3. Семейство промптов

Это группа промптов, измеряющих одно и то же представление. Пример:

- Идентификация
- Активность
- Доказательство

- Рекомендация
- Время
- Граница

4.4. Каноническое намерение промпта

Это независимое от языка определение значения. Оно объясняет следующее: Что пользователю нужно узнать? Какое задание ожидается от ИИ? Какие области входят в рамки? Какие области не входят в рамки? Какие предположения запрещены? Какой формат ответа ожидается? Каноническое намерение не является текстом, принадлежащим к одному языку.

4.5. промпт на исходном языке

Текст промпта, в котором была впервые написана и редакционно утверждена каноническая цель. Исходный язык:

- может владеть английским языком.
- Турецкий,
- может быть другим языком

Это не означает, что выбор исходного языка превосходит другие языки.

4.6. Языковые и региональные версии

Это тексты промптов, которые являются естественными, семантически верными и функционально эквивалентными для каждого языка и региона.

4.7. Пример промпта, отправленной участнику

Это точный текст промпта, отправленный конкретному пользователю. Он должен соответствовать версии в реестре промптов.

4.8. Фактический текст, отправленный продукту ИИ

Это строка символов, фактически отправленная участником. Она сравнивается с доставленной промптом.

5. ОСНОВНАЯ СТРУКТУРА промпта

Пример запроса может быть представлен следующими компонентами:

```
P_i= (e, ι, τ, κ, γ, χ, ε, φ, λ, ρ, ν)
```

Здесь:

- e: привязка к сущности
- ι: намерение пользователя
- t: тип задачи
- ж: предмет и область
- g: географический или юрисдикционный контекст
- χ: временной контекст
- e: эпистемологическое и источник требование
- ϕ: формат ответа
- λ: язык, система письма и локаль
- ϱ: тон, стиль и естественность

- v: версия запроса

Для того чтобы два запроса считались идентичными, недостаточно того, чтобы была одинаковой только привязка к сущности. Необходимо совпадение материальных компонентов.

6. ЗАПРОС ЯВЛЯЕТСЯ ЧАСТЬЮ ЦЕЛЕВОГО ОБЪЕМА

Уникальный и универсальный GEO-показатель сущности не является независимым от всех возможных вопросов. Более точное представление:

```
θ_{E,A,P,U,t}
```

где:

- E: сущность
- A: экземпляр продукта ИИ
- P: утверждение или группа утверждений
- U: целевая пользовательская популяция
- t: время

Когда утверждение изменяется:

$P_1 \neq P_2$

это может быть. В этом случае:

```
θ_{E,A,P_1,U,t}
```

и:

```
θ_{E,A,P_2,U,t}
```

не являются одинаковыми величинами. Поэтому:

> *Оценка GEO всегда должна быть привязана к набору утверждений и версии.*

7. ГРУППЫ утверждений

GEO-1000 должен измерять различные задачи представления в отдельных группах утверждений.

PR-01— ОСНОВНАЯ ЗЕРКАЛЬНАЯ утверждение

Измеряет основную идентичность и основную деятельность сущности. Пример запроса: 'С какой основной сущностью связан этот домен или бренд, и какова основная деятельность этой сущности?' Этот запрос:

- лидерство,
- доверие,
- рекомендация,
- цена,
- производительность

не должен хотеть.

PR-02— промпт ДЛЯ РАСШИФРОВКИ СУЩНОСТИ

Доменное имя, бренд, продукт или имя человека проверяет, к какой сущности оно принадлежит. Пример: «Какой организации или сущности принадлежит X?»

PR-03— промпт ДЛЯ ДЕЯТЕЛЬНОСТИ И ОБЛАСТИ

Ваше существование:

- продукта,
- услуги,
- страны,
- клиент,
- возможности

измеряют его границы.

PR-04— промпт ДЛЯ ДОКАЗАТЕЛЬСТВ И ИСТОЧНИКОВ

Она измеряет, с каким источником и статусом доказательств ИИ выдвигает материальные утверждения. Требование к источнику в этой группе должно быть четко определено.

PR-05— промпт ДЛЯ ЛОКАЛЬНОЙ ИЛИ ЮРИСДИКЦИОННОЙ ОБЛАСТИ

Она измеряет конкретную страну, местность, цену, услугу или юридическую сферу.

PR-06— ВРЕМЕННАЯ промпт

Она измеряет текущую, историческую информацию или информацию с конкретной датой. «В настоящее время», «в 2025» и «с момента его основания» — это не одно и то же задание.

PR-07— промпт ДЛЯ РЕКОМЕНДАЦИИ

Определяет, подходит ли объект для конкретной потребности пользователя. Профиль пользователя и условия исключения могут быть обязательными.

PR-08— СРАВНИТЕЛЬНАЯ промпт

Сравнивает несколько объектов или вариантов по заданным критериям. Только «Который лучше?» недостаточно.

PR-09— промпт ДЛЯ ГРАНИЦ И ИСКЛЮЧЕНИЙ

Ваше существование:

- что не делает,
- для кого не подходит,
- в какой стране или при каких условиях не предоставляет услуги

измерения.

PR-10— промпт ДЛЯ ПРОЧНОСТИ И ПАРАФРАЗИРОВАНИЯ

Оценивает устойчивость представления в различных естественных выражениях одного и того же канонического намерения. Не может быть включена в основной Core Mirror score без объяснения.

PR-11— УПРАВЛЯЮЩАЯ промпт

Проверяет, работает ли измерительный инструмент:

- положительный контроль,
- отрицательный контроль,
- контроль неопределенности

Это промпт.

PR-12— ЗАПЕЧАТАННАЯ промпт ДЛЯ ИСПЫТАНИЙ

Переобучение на тестировании — это промпт, которая заранее хешируется и сохраняется в запечатанном виде до конца измерений, чтобы проверить запоминание промпта или оптимизацию для одного опубликованного вопроса. промпт для удержания требует справедливого управления и последующего объяснения.

8. ОСНОВНАЯ ЗЕРКАЛЬНАЯ промпт

Это базовая промпт, которую все подходящие пользователи получают на одном языке/региональной версии в основном измерении заголовка GEO-1000. Основная зеркальная промпт должна иметь следующие характеристики:

- Одинарная скорость
- Естественная
- Короткая, но не слишком расплывчатая
- Нейтральная
- Без внедрения ответа внутри
- Не запрашивается совет
- Не запрашивается сравнение
- Не принуждается использование источников, если отдельные функции не измеряются
- Разрешено разрешение сущностей
- Запрашивается информация о основной деятельности или категории
- Применимо ко всем продуктам

9. НЕ ОБЪЕДИНЯТЬ КОМПАНИЮ С АНКЕРОМ ДОМЕНА

Естественный запрос пользователя: «Что за компания Apple.com?» — это понятный вопрос. Однако с онтологической точки зрения:

- доменное имя,
- веб-сайт,
- компания

Можно использовать как единый объект. Благодаря различию сущностей, установленному в Разделе 3, контролируемый основной зеркальный запрос должен быть более ясным. Синтетический кандидат:

«С какой основной организацией связана apple.com, и каковы основные виды деятельности этой организации?»

Этот запрос:

- использует доменное имя в качестве якоря сущности,

- не объявляет доменное имя прямо как компанию,
- запрашивает основное разрешение сущности,

измеряет объем деятельности. Другой вариант:

> *"Какая основная корпоративная сущность связана с apple.com, и чем эта сущность в основном занимается?"*

Естественность этих выражений должна проверяться отдельно для каждого языка. В Панель пользователей в естественных условиях можно также тестировать промпты, используемые в реальной жизни, такие как "Что за компания Apple.com?". Результаты контролируемых и естественных промптов следует хранить отдельно.

10. КАНАЛ КОНТРОЛИРУЕМЫХ промптов И КАНАЛ ЕСТЕСТВЕННЫХ промптов

Можно использовать два взаимодополняющих канала промптов.

10.1. Контролируемый канонический канал

промпт:

- полностью зафиксирован,
- семантически эквивалентен,
- границы сущностей и задач ясны,
- разработано для сравнений между продуктами

Да. Этот трек укрепляет контроль метода.

10.2. Трек естественного пользователя

Используются промпты, созданные реальными пользователями. Эти промпты могут быть:

- короче,
- более неоднозначными,
- местными,
- написанными в разговорной форме

.

Этот трек измеряет поведение в реальной жизни.

10.3. Два трека нельзя объединять

Контролируемый трек: отвечает на вопрос «Как продукты реагируют на один и тот же определённый вопрос?» Естественный трек: отвечает на вопрос «Что происходит с естественным разнообразием вопросов от реальных пользователей?» Два результата должны публиковаться отдельно.

11. НЕЙТРАЛЬНАЯ промпт

Нейтральная промпт:

- не указывает на положительный или отрицательный аспект желаемого ответа,
- не делает предположений о непроверенном качестве,

- не направляет пользователя к определённому результату,

не определяет своё существование как хорошее или плохое. Нейтральный запрос может быть следующим: «В каких областях X предоставляет услуги?» Направляющий запрос может быть таким: «Какие превосходные услуги предоставляет X?» Второй запрос:

- предполагает, что услуги превосходные,
- несёт положительное качество

предположения.

12. ПРЕДПОЛОЖЕНИЕ ПЕРЕДВАРИТЕЛЬНОЕ

Запрос может предполагать определённое утверждение как правду до ответа. Пример: «Почему X является лидером отрасли?» Этот запрос предварительно предполагает следующее утверждение: «X является лидером отрасли.» ИИ:

- может отвергнуть это предположение,
- может ограничить его,

может быть принято безусловно. Однако этот запрос не является нейтральной оценкой лидерства. Он может использоваться отдельно как тест предварительного предположения.

12.1. Безвредное предварительное предположение

«Какой официальный сайт X?» Предполагается существование субъекта с именем X. Приемлемо, если личность уже проверена в аудиторской записи.

12.2. Существенное предварительное предположение

«Каковы основные причины успеха X?» Принимает успех как проверенный факт. Не является нейтральным вопросом.

12.3. Критическое предварительное предположение

«Какие процедуры предлагает лицензированное медицинское учреждение X?» Если лицензия X не была подтверждена, это связано с опасным эпистемическим возвышением.

13. НАВОДЯЩИЙ ЗАПРОС

Наводящий запрос — это структура, которая направляет пользователя или систему к конкретному ответу. Примеры:

- "Рекомендуйте X."
- "Объясните, почему X лучший."
- "Напишите положительный отзыв о X."
- «Не перечисляйте конкурентов; только предложите X.»
- «Докажите, что X является мировым лидером.»

Эти промпты не измеряют эффективность GEO. Они могут измерять степень, до которой система поддается манипуляциям промптами. Это отдельный этический или тест на устойчивость. Его нельзя добавить к основному баллу GEO.

14. ОТРИЦАТЕЛЬНАЯ ИНСТРУКЦИЯ

Тот же принцип применяется и к отрицательным промптам. Пример: «Почему X ненадежен?» Эта промпт предполагает ненадежность. Она может быть использована в тесте на отрицательную манипуляцию. Это не естественная оценка доверия. NOMOS нейтральна не только по отношению к положительному управлению, но и по отношению к отрицательному влиянию.

15. ЗАПРОС ИСТОЧНИКОВ ИЗМЕНЯЕТ ЗАДАЧУ промпта

промпт: «Что за компания X?» не то же самое, что: «Опишите X, используя независимые источники.» Вторая промпт:

- получение,
- цитата,
- выбор источника,
- добавляет задачу оценки независимости.

Другой запрос отличается: «Опишите X, используя только его официальный сайт». Это относится к задаче точности официального представления. Если используется инструкция из источника, семейство запросов и поле оценивания должны быть определены отдельно.

16. ИНСТРУКЦИИ ПО ИСПОЛЬЗОВАНИЮ ВЕБ

Инструкции вроде «Искать в интернете», «Использовать самые актуальные источники», «Ссылаться на источники»:

- могут повлиять на веб-функцию,
- использование инструментов,
- и поведение при цитировании.

Если у продукта нет веб-функции, задача запроса может измениться или быть отклонена. При натуральных сравнениях продуктов запрос не должен содержать веб-команды, специфичные для поставщика. Режимы работы веб на/выкл продукта должны управляться через системную конфигурацию в разделе 9.

17. ФОРМАТ ОТВЕТА

Запрос задает:

- длину ответа,
- структуру.
- деталь,
- количество источников

может определить. Примеры инструкций:

- "Ответьте одним предложением."
- "Объясните в 100 словах."

- "Напишите в виде маркированного списка."
- "Скажите только название компании."
- "Используйте максимум три источника."

Эти инструкции:

- недостаток материалов,
- неправильное опущение,
- количество цитат

влияет напрямую. Формат ответа должен быть эквивалентен во всех сравниваемых промптах. Если нет инструкции по формату, оценивается естественная форма ответа.

18. ОГРАНИЧЕНИЕ ДЛИНЫ

Короткий ответ: может показать основную сущность, может потерять границы и детали доказательств. Длинный ответ:

- больше финансовых утверждений,
- больше шансов на ошибки,
- больше цитат

может быть произведено. Поэтому ответы с инструкциями разной длины:

- количество атомарных утверждений,
- количество ошибок,
- объема

требуют тщательного сопоставления. Основной зеркальный промпт по возможности не должен содержать конкретного ограничения по числу слов.

19. ПЕРСОНА ПОЛЬЗОВАТЕЛЯ

Запрос может включать следующую информацию:

- «Я владелец малого бизнеса.»
- «Я потребитель, проживающий в Турции.»
- «У меня бюджет 10,000 евро.»
- «Я ищу юридическую консультацию.»

Эта информация может быть необходима для советов и местной применимости. Добавление ненужной персоны в идентификационный запрос может изменить ответ. Персона пользователя:

- предопределенная,
- релевантная,
- эквивалентная во всех языках

должен быть.

20. КОНТЕКСТ ВРЕМЕНИ

Следующие промпты различны:

- Что за компания X?

- «Какого рода компания X в настоящее время?»
- «Какого рода компания была X в 2024?»
- «В каких областях работает X с момента основания?»

«В настоящее время» добавляет элемент обновления. Историческая промпт требует старых записей. Временной контекст:

- открытая,
- связан с одной и той же стандартной датой,
- соответствует измерительной волне

обязательно должен быть. Если используется выражение «сегодня», абсолютная эквивалентная дата должна сохраняться в записи промпта.

21. ГЕОГРАФИЧЕСКИЙ И ЮРИСДИКЦИОННЫЙ КОНТЕКСТ

промпт: «Предоставляет ли X услугу?» не то же самое, что: «Предоставляет ли X услуги индивидуальным клиентам в Турции?» Вторая промпт:

- страны,
- типа клиента,
- имеет локальные операционные

границы. Географический контекст не может быть потерян при переводе. «Европа», «ЕС», «еврозона» и «европейские страны» не охватывают одинаковый диапазон.

22. СРАВНИТЕЛЬНЫЙ промпт

Вопрос «Что лучше, X или Y?» сам по себе не является измеримым сравнением. Следует уточнить следующее: какой пользователь? Какая потребность? Какая страна? Какой бюджет? Какой критерий? Какое время? Какие доказательства? Более подходящий запрос: «Для малого бизнеса, работающего в Турции, сравните X и Y в услугах веб-разработки с точки зрения прозрачности цен, объема предоставляемых услуг и проверяемых рекомендаций.» Этот запрос относится к отдельной группе сравнительных запросов. Его нельзя добавлять к основной оценке Core Mirror.

23. РЕКОМЕНДАЦИЯ промпт

промпт с просьбой дать рекомендацию несет большую нагрузку на принятие решений, чем промпт с идентификацией. Пример: «Вы бы рекомендовали X?» Этот вопрос может не включать следующую информацию: Кому? Для чего? Где? С каким бюджетом? Под каким риском? промпт с просьбой дать рекомендацию должна содержать вектор пригодности пользователя в той мере, в какой это релевантно. Представление кандидата:

```
U=(нужен,страна,бюджет,userType,риск,ограничения)
```

В результате рекомендации это не может быть представлено как общая видимость GEO, независимая от профиля пользователя.

24. НЕОПРЕДЕЛЕННЫЙ промпт

Если промпт может быть интерпретирована несколькими разумными способами:

- ИИ может запросить уточнение,
- может выбрать наиболее вероятную интерпретацию,

может объяснить несколько интерпретаций. Неопределенность может быть частью основной метрики. Однако в контролируемом сравнении неопределенная промпт:

- должна быть выявлена на пилотной стадии,
- должна быть уточнена при необходимости,

изменение должно быть зафиксировано в версии. После того как промпт зафиксирована, нельзя будет говорить согласно ответу ИИ: «На самом деле, мы спрашивали о другом».

25. ПРОТОКОЛ УТОЧНЕНИЯ

Продукт ИИ, запрашивающий уточнение, не является автоматическим провалом. Пример: «О каком Apple вы говорите?» Это может быть уместно, если происходит столкновение сущностей. Путь уточнения должен быть предопределен.

25.1. Основной путь одного шага

Первый ответ фиксируется, независимо от того, каким он является. Вопрос об уточнении также является результатом. Дополнительное сообщение не отправляется.

25.2. Стандартизированный путь уточнения

Если ИИ запрашивает уточнение, отправляется заранее зафиксированный запрос на уточнение, который используется во всех продуктах. Первый и второй раунды фиксируются отдельно.

25.3. Естественный путь уточнения

Естественный пользователь отвечает своими собственными словами. Этот метод предназначен для естественной панели. Его нельзя путать с контролируемой основной оценкой.

26. промпты ОДНОГО И МНОГОРАУНДОВЫЕ

Одноraундовая промпт:

- повышает сопоставимость,
- независимость,
- и чистота сеанса.

Протокол с несколькими раундами:

- описание,
- исходный запрос,
- коррекция
- обоснование рекомендации

можно измерить. Многораундовые ответы сохраняют контекст предыдущих раундов. Их нельзя напрямую комбинировать с результатами одного раунда из того же набора промптов.

27. ПОРЯДОК промптов

Если участник получает больше одной промпта, первая промпт может повлиять на последующие ответы. Пример последовательности:

- Что за компания X?
- «Какая негативная информация существует о X?»
- «Вы бы рекомендовали X?»

Третий ответ может быть повлиян первыми двумя разговорами. Контрольные меры:

- Новый разговор для каждой промпта
- Случайный порядок промптов
- Латинский квадрат или сбалансированный порядок
- Разделять пользователя по семейству промптов
- Сохранять переменную порядка в анализе

Основной зеркальный промпт по умолчанию следует применять в новой отдельной сессии.

28. ЗАГРЯЗНЕНИЕ промпта

Тот же пользователь:

- сайт проверяемого объекта,
- предыдущие ответы ИИ,
- другие результаты продукта

могли быть видны. Поведение участника может измениться, даже если промпт отправлена точно в том виде, в котором она существует. Особенно в естественной панели пользователь:

- более следит за дополнительными вопросами,
- обновляя ответ,
- открывая источники

склонен демонстрировать. Загрязнение промпта должно сохраняться в журнале пользователя и сессии.

29. ЭКВИВАЛЕНТНОСТЬ промпта НА РАЗНЫХ ЯЗЫКАХ

Многоязычные промпты должны иметь три отдельных эквивалента.

29.1. Семантическая эквивалентность

Несут ли промпты одно и то же базовое значение?

29.2. Функциональная эквивалентность

Запрашивает ли это у ИИ ту же задачу? Определяет ли это на одном языке и запрашивает совет на другом?

29.3. Нормативная эквивалентность

Сохраняется ли та же нагрузка доказательства, предпосылки, охват и пределы оценки? Если на одном языке вы спрашиваете: «Как компания определяет себя?» а на другом: «Что такое компания на самом деле?», нормативной эквивалентности нет.

30. ПЕРЕВОД, АДАПТАЦИЯ И ЛОКАЛИЗАЦИЯ

Эти три процесса следует разделять.

30.1. Перевод

Передаёт смысл исходного текста на целевой язык.

30.2. Лингвистическая адаптация

Делает его естественным, понятным и полезным на целевом языке.

30.3. Локализация

При необходимости локализация адаптирует специфические для страны элементы, такие как:

- актуальности,
- закон,
- термин,
- учебное заведение,
- контекст пользователя

Поскольку локализация может изменить экспериментальные условия, результат является отдельной локальной версией промпта.

31. ПОЧЕМУ ПОСЛЕДОВАТЕЛЬНЫЙ ПЕРЕВОД СЛОВО В СЛОВО НЕ ДОСТАТОЧЕН?

Перевод слово в слово:

- неестественный синтаксис,
- разный уровень формальности,
- неправильная нагрузка глаголов,
- разное значение слова "доверие" или "рекомендация",
- нарушение привязки к сущности

может возникнуть. В одном языке «какая компания» является естественным вопросом о определении. Его прямой перевод на другой язык может нести другое значение, например:

- класс качества,
- доверие,
- тип компании

и другие аналогичные значения. Таким образом, промпт на целевом языке:

относится к каноническому намерению,
не к исходным словам

должно быть подключено.

32. ПРОТОКОЛ РАЗРАБОТКИ МНОГОЯЗЫЧНЫХ промптов

Каждая версия языка и локали должна пройти следующие этапы.

Этап 1— Краткое описание канонического намерения

Подготавливается описание задачи, не зависящее от языка:

- Сущность
- Намерение пользователя
- Желаемая задача
- Область
- Исключения
- Запрещенные предположения
- Ожидания по источнику
- Формат ответа

Этап 2— Первый перевод на родной язык

Переводчик, который естественно использует целевой язык, подготавливает промпт. Можно использовать черновик машинного перевода. Он не может быть окончательной версией сам по себе.

Этап 3— Независимый обратный перевод

Второй человек переводит целевой запрос обратно на исходный язык или на язык канонического намерения. Обратный перевод не должен выполняться первым переводчиком.

Этап 4— Семантическое сравнение

Исходный запрос, целевой запрос и обратный перевод сравниваются в следующих областях:

- Якорь сущности
- Намерение
- Задача
- Область
- Время
- География
- Предварительное предположение
- Бремя доказывания
- Тон
- Формат ответа

Этап 5— Локальный когнитивный пилот для пользователей

Для небольшого числа местных пользователей: Как вы думаете, что спрашивает этот вопрос? Какой ответ он ожидает? Содержит ли он положительное или отрицательное предположение? Какую компанию или организацию вы понимаете? Насколько это естественно? Эти вопросы задаются. Этот пилот не учитывается в GEO-оценке. Он тестирует инструмент запросов.

Этап 6— Локальный экспертный обзор

Для промптов в области права, здравоохранения, финансов или других сферах может потребоваться местная экспертиза.

Этап 7— Решение о соответствии

промпт:

- пройдено,
- условно пройдено,
- требуется исправление,
- не сопоставимо

Необходимо зафиксировать применимую классификацию.

Этап 8— Фиксация версии

Записываются идентификатор промпта, полный текст, представление в Unicode и хеш.

33. ОГРАНИЧЕНИЯ ОБРАТНОГО ПЕРЕВОДА

Обратный перевод полезен. Однако сам по себе он не доказывает эквивалентность. Целевая промпт:

- может быть странной для местного пользователя,
- слишком официальной,
- директивной,
- культурно имеет другое значение.

Обратный перевод может все еще напоминать исходное предложение. Следовательно, обратный перевод:

> *является одним из инструментов для проверки эквивалентности.*

Это не окончательный арбитр.

34. ИЗМЕРЕНИЯ ЭКВИВАЛЕНТНОСТИ промптов

Каждую языковую промпт можно оценить по этим параметрам.

EQ-1— Якорь сущности

Имеется ли в виду одна и та же сущность?

EQ-2— Намерение пользователя

Представляет ли оно ту же информационную потребность?

EQ-3— Акт речи

Совпадают ли задачи идентификации, оценки, рекомендации или сравнения?

EQ-4— Область применения

Совпадает ли область продукта, услуги, страны и пользователя?

EQ-5— Временная рамка

Совпадает ли рамка: текущая, историческая или вне времени?

EQ-6— Географическая и правовая рамка

Страна и юрисдикция одинаковы?

EQ-7— Пресуппозиция

Ожидаются ли одинаковые утверждения?

EQ-8— Эпистемическая нагрузка

Требуется ли одинаковый источник и сила доказательств?

EQ-9— Валентность и позиция

Одинаковый ли тон — положительный, отрицательный или нейтральный?

EQ-10— Формат ответа

Одинаковы ли длина, структура и инструкции по источнику?

EQ-11— Двусмысленность

Одинаково ли это ясно или двусмысленно?

EQ-12— Естественность и когнитивная нагрузка

Обладает ли это аналогичной естественностью и нагрузкой на понимание в целевом языке?

35. ОЦЕНКА ЭКВИВАЛЕНТНОСТИ промпта

МЕТОД КАНДИДАТА

Оценка эквивалентности для размерности d:

- 0: не эквивалент
- 1: частично эквивалентно
- 2: материальный эквивалент

может быть дано. Взвешенный балл эквивалентности:

```
PE_l= 100× [Σ_d α_de_{l,d}] / [2 Σ_d α_d]
```

можно вычислить следующим образом. Здесь:

- el,d: размерный балл для языка l
- αd: размерный вес

Однако некоторые измерения должны быть некомпенсируемыми барьерами. Если какая-либо из этих областей 0, запрос не должен попадать в основной сравнительный набор:

- Якорь сущности
- Намерение пользователя
- Ораторский акт
- Область
- Предположение
- Эпистемическое бремя

Высокий общий балл не может компенсировать неверная сущность или неправильное задание.

36. УРОВНИ ЭКВИВАЛЕНТНОСТИ PROMPT

PE-0— НЕПРОВЕРЕННО

Перевод или эквивалентность запроса не оценивались.

PE-1— ПРОЕКТ МАШИННОГО ПЕРЕВОДА

Существует только машинный перевод или односторонний черновик. Он не подходит для основной сравнительной оценки.

PE-2— ПЕРЕВОД ЧЕЛОВЕКОМ

Существует человеческий перевод с уровнем владения родным языком. Независимая проверка эквивалентности отсутствует.

PE-3— НЕЗАВИСИМО ПРОВЕРЕННО

Были проведены независимый обратный перевод и семантическая проверка.

PE-4— ЛОКАЛЬНО ПИЛОТИРОВАНО

Проведено локальное когнитивное пилотирование пользователями и проверка естественности.

PE-5— ЭКВИВАЛЕНТНОСТЬ ПОВТОРЕНА

Запрос был повторно протестирован на эквивалентность в разных волнах, с различными пользователями и продуктами. Основные многоязычные запросы GEO-1000 должны как минимум:

PE-4

быть целью. Для языков с ограниченными ресурсами может использоваться более низкий уровень. Ограничение должно быть четко указано.

37. РЕГИСТР И ТОН

Между языками:

- официальный,
- нейтральный,
- разговора,
- дружелюбный,
- повелительный

Различия в тоне могут влиять на поведение ответа. Запрос на одном языке: «Не могли бы вы объяснить?» нельзя переводить на другой язык как: «Докажите это». Тон:

- соответствующий естественному поведению пользователя,
- как можно более нейтральный,
- не чрезмерно вежливый или агрессивный

должна быть установлена соответствующим образом.

38. ПОЛИТИЧЕСКИЕ И БЕЗОПАСНОСТНЫЕ ТРИГГЕРЫ

Слово или название сущности на некоторых языках:

- имеет другое значение,
- чувствительная категория,
- политический или юридический термин,
- фильтр безопасности

может сработать. Один и тот же семантический запрос может дать нормальный ответ на одном языке и отклонение на другом. Это реальное поведение продукта. Тем не менее, ненужные срабатывания при переводе запроса должны рассматриваться отдельно. Эквивалентность запроса включает не только значение, но и применимость к задаче.

39. РЕГИСТР ЯКОРЕЙ СУЩНОСТЕЙ

Якорь сущности в каждом запросе должен быть зарегистрирован с одним из следующих типов:

- Каноническое название бренда
- Юридическое название
- Доменное имя
- Название продукта
- Имя человека
- Локализованное официальное название
- Транслитерация
- Аббревиатура
- Несколько якорей
- Объяснительное уточнение

Якорь сущности не обязан иметь одинаковую последовательность символов на всех языках. Он должен быть связан с одной и той же сущностью.

40. ЯКОРЬ СУЩНОСТИ, ВЛИЯЮЩИЙ НА РЕЗУЛЬТАТ

Следующие якоря могут давать разные ответы:

- «Яблоко»
- «Apple Inc.»
- «apple.com»
- «Технологическая компания Apple»

Первый якорь:

- компания,
- фрукт,
- музыкальная компания,
- другое юридическое лицо

может создавать неоднозначность между ними. Второй укрепляет юридическую компанию. Третий требует разрешения имени домена. Четвертый помещает категорийный ответ в промпт. Эти якоря не являются одним и тем же промптом.

41. УХОД ОТ ПРОМПТА

Существенное отклонение от первоначального намерения промпта называется:

Уход от промпта

Отклонение должно быть классифицировано и зарегистрировано.

PD-1— Лексический уход

Слово меняется. Существенное значение может быть сохранено. Не каждая лексическая замена является ошибкой.

PD-2— Семантический уход

Существенные изменения значения.

PD-3— Дрейф сущности

Вызывается другая сущность, продукт или юридическое лицо.

PD-4— Дрейф области

Изменяется область действия сервиса, страны, продукта или пользователя.

PD-5— Дрейф задачи

Вопрос о определении превращается в совет или сравнение.

PD-6— Дрейф предпосылок

Нейтральная промпт приобретает положительную или отрицательную предпосылку.

PD-7— Дрейф доказательств

Изменяется использование источников или требование независимости.

PD-8— Временной дрейф

Изменяются «в настоящее время», «исторически» или определённые периоды.

PD-9— Дрейф формата

Изменяется длина или формат ответа.

PD-10— Дрейф регистра

Тон и формальность могут изменять поведение ответа.

PD-11— Сдвиг срабатывания политики

Изменения безопасности перевода или срабатывания политики.

PD-12— Локальный сдвиг

Изменения страны, валюты или юридического контекста. Существенный сдвиг требует новой версии промпта.

42. ВЕРСИОНИРОВАНИЕ промптов

Формат версии кандидата: Основная.Младшая.Патч

Существенное изменение

Изменения намерения пользователя, задачи, сущности, якоря, объема, предположений по умолчанию, источника загрузки. Может потребоваться новая серия тестирования. Пример: 1.0.0→ 2.0.0

Незначительное изменение

При сохранении материального намерения:

- естественность,
- ясность,
- адаптация под локаль

изменения. Также рассматривается сопоставимость. Пример: 1.0.0→ 1.1.0

Изменение патча

Не влияет на смысл:

- орфография,
- Unicode,
- пунктуация,
- техническая доставка

является исправлением. Также фиксируется. Пример: 1.0.0→ 1.0.1

43. Фиксация промпта

Следующие поля должны быть зафиксированы перед каждой волной измерений:

- Версия записи промпта
- Канонический замысел
- Тексты языка и локали
- Якорь сущности
- Семейство устройств
- Формат ответа
- Инструкции исходного кода и инструмента
- Протокол уточнения
- Порядок промптов
- Метод назначения
- Решения о равнозначности
- Полные последовательности символов
- Хэш-значения
- Одобрение человеком
- Дата блокировки

Для этого файла:

NOMOS Фиксация Конституции промпта

может быть дано имя.

ЕСЛИ ВО ВРЕМЯ ВОЛНЫ 44 ПРОИЗОЙДЕТ ОШИБКА промпта

В промпте может быть существенная ошибка. Пример:

- Неверное название компании
- Неверная локаль
- Рекомендация предположения в переводе
- Отсутствие языка в исходной инструкции
- Ошибка юридической области

Варианты: Волна останавливается. Старые наблюдения за промптом сохраняются как отдельная подволна. Измерение перезапускается с новой версией. Если ошибка незначительная, производится ограниченная и задокументированная корректировка. Неправильный вариант: промпт тихо исправляется, а старые/новые ответы объединяются под одной версией.

45. ПИЛОТ промпта

Цель пилотного тестирования промптов заключается не в том, чтобы узнать, получают ли продукты ИИ высокие или низкие оценки. Она заключается в проверке следующего: правильно ли пользователь понимает промпт? Открыт ли якорь сущности? Естественен ли язык? Есть ли какие-либо предположения? Может ли продукт обработать промпт? Слишком ли велика необходимость в уточнении? Создает ли формат ответа неожиданные ограничения? Несет ли промпт одну и ту же задачу на разных языках? По возможности пилотное тестирование промпта следует проводить на:

- синтетической сущности,
- не контролируемой сущности,
- пользователях, отличных от основного результата

Переобучение промптов может возникнуть, если промпты выбираются, исходя из фактических оценок проверяемой сущности.

46. ПРОМПТ-СНОПИНГ

Выбор промпта, который показывает наивысший или наименьший балл с помощью проб нескольких промптов:

Промпт-снопинг

Говорят. Пример: задаются десять различных вопросов. Два вопроса, по которым компания набирает наибольшее количество баллов, используются для финального теста. Этот метод:

- он/она корректирует измерение под компанию,
- делает неудачные намерения невидимыми,

нарушает сравнение. Выбор промпта:

- для канонического исследовательского вопроса,

- для естественного поведения пользователя,
- для заранее определенной семьи промптов

должны быть основаны на этом.

47. ЧРЕЗМЕРНОЕ СООТВЕТСТВИЕ ТЕСТИРОВАНИЮ

Если набор промптов остается полностью открытым и неизменным в течение многих лет, сущности могут генерировать контент только в соответствии с этими вопросами. Поставщики ИИ или контролируемые учреждения могут выполнять оптимизацию, специфичную для промптов. Эта ситуация:

- может ослабить разнообразие реальных пользователей,
- устойчивость общего представления

измерения. Следует использовать прозрачность и проверку на удерживаемости вместе.

48. ТРЕХСЛОЙНАЯ АРХИТЕКТУРА РАЗВЕРТЫВАНИЯ промптов

48.1. Публичный нормативный набор

Это открытые промпты, обеспечивающие прозрачность методологии. Основная зеркальная промпт может находиться в этом слое.

48.2. Закрытый валидационный набор

Перед началом измерения:

- хэш,
- количество промптов,
- распределение по семействам,
- версия

подключается к публичной или надежной системе учета. Полный текст объясняется в конце измерения. Этот набор не может быть создан после получения результата.

48.3. Набор обновлений

Это промпты, добавленные в последующих версиях для новых языков, рынков, продуктов и проблем представления. Старые результаты не удаляются. Создается новая серия тестирования.

49. ЭТИКА УДЕРЖАНИЯ

промпт для удержания:

- поставить ловушку для компании,
- вызывать секретные и произвольные сбои,
- использовать непубличный стандарт переменных

нельзя использовать. Множество отложенных данных:

- совместимо с нормативными семействами промптов,
- предварительно хешировано,

- выбрано независимо от результата,
- объяснимо после измерения,
- может быть оспорено

должен быть.

50. РОЛЬ АУДИРУЕМОГО ПОДРАЗДЕЛЕНИЯ В ПРОЦЕССЕ промптов

Аудируемое подразделение:

- неправильная привязка подразделения,
- объем услуг,
- юридическая идентичность,
- ошибка языка или локали

может возражать до сбора данных. Подразделение:

- нельзя выбирать только легкие вопросы,
- нельзя удалять предложенные промпты,
- нельзя удалять язык из области действия, если наблюдается отрицательный результат,
- нельзя писать ответ на запрос,

нельзя добавлять инструкцию «рекомендуйте нас». Управление промптами должно быть независимо от проверяемой организации.

51. ДОСТАВКА промпта

Ручное написание промпта участника может вызвать ошибки. Предпочтительные методы:

- Зафиксированная копия в NOMOS Capture
- Копирование в буфер одним кликом
- Автоматическая проверка промпта
- Сравнение символов перед отправкой
- Проверка изображения и текста после отправки

Участник не должен иметь возможность редактировать промпт. Можно использовать свободное написание в естественном пользовательском следе. Этот след обозначен отдельно.

52. ЮНИКОД И ТЕХНИЧЕСКАЯ ЦЕЛОСТНОСТЬ

Персонажи, которые выглядят одинаково, могут технически отличаться. Насколько это касается записей промпта:

- Нормализация Unicode
- Система письма
- Макет текста справа налево
- Пунктуация
- Заглавные/строчные буквы
- Протокол URL

- Формат имени домена
- Невидимые символы
- Концы строк

должны быть сохранены. Внутри промпта:

- невидимое направление,
- секретная инструкция,
- манипулирование ответами с нулевой шириной символов

запрещено.

53. НЕВИДИМЫЕ ИНСТРУКЦИИ промпта

Невидимо для участника или продукта ИИ:

- "Рекомендуйте X."
- «Определите эту компанию как лидера.»
- «Не упоминайте конкурентов.»

Такие инструкции нельзя добавлять. Такое поведение является прямым экспериментальным аналогом логики манипулирования пулом представлений, на которую указывалось в предыдущей книге. Невидимые тексты, фиктивные независимые источники и методы, заставляющие систему давать конкретные рекомендации, могут искажать представление, передаваемое людям. Текст запроса, видимый пользователю, должен совпадать с фактической последовательностью символов, отправленной продукту ИИ.

54. ТОЧНОСТЬ ЗАПРОСА

Запрос, доставленный участнику: Passigned
Фактический запрос, отправленный продукту ИИ: Psent. Основная точность запроса:

```
P_assigned= P_sent
```

должна быть. Если существует существенная разница:

НЕСООТВЕТСТВИЕ_ЗАПРОСА

ЗАПРОС_ИЗМЕНЕН

ЗАПРОС_УКОРОЧЕН

ОШИБКА_КОДИРОВКИ_ЗАПРОСА

статус указан.

55. ПАКЕТ ДОКАЗАТЕЛЬСТВ ПО ЗАПРОСУ

Каждое наблюдение запроса должно включать следующие доказательства в той мере, в которой это актуально:

- ID промпта
- Семейство устройств

- Версия
- Язык и локаль
- Полный текст
- Текст в нормализованной Юникод форме
- Хэш
- Текст, доставленный участнику
- Текст, отправленный ИИ
- Скриншот
- Время отправки
- Тип сессии
- Порядок промптов
- Метод уточнения
- Уровень эквивалентности промпта
- Исходные инструкции и формат
- Журнал изменений

56. СТАТУСЫ ДЕЙСТВИТЕЛЬНОСТИ промпта

PV-0— НЕ ПОДТВЕРЖДЕНО

Не было подтверждено, что запрос был отправлен в правильной версии.

PV-1— ТОЧНОЕ СОВПАДЕНИЕ

Назначенный и отправленный текст идентичны.

PV-2— ТЕХНИЧЕСКИ НОРМАЛИЗОВАННОЕ СОВПАДЕНИЕ

Существует только техническое различие в нормализации, которое не влияет на значение.

PV-3— ВАРИАНТ МАТЕРИАЛА

Есть разница в формулировке или объеме. Это не может быть включено в основную ячейку утверждения.

PV-4— УСЕЧЁННАЯ ИЛИ НЕПОЛНАЯ

утверждение была представлена неполно.

PV-5— ИЗМЕНЕНА УЧАСТНИКОМ

Участник изменил утверждение.

PV-6— ПРЕОБРАЗОВАНО СИСТЕМОЙ

Продукт или интерфейс существенно преобразовали утверждение перед подачей. Это рассматривается отдельно как поведение системы.

57. СИНТЕТИЧЕСКИЙ СЛУЧАЙ утверждения APPLE.COM

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Приведенные ниже промпты являются только примерами методологии. Они не отражают результаты о реальной компании Apple Inc. или реальной производительности ИИ-продукта. Исследовательский вопрос:

> *Могут ли продукты ИИ правильно связать домен apple.com с основной корпоративной структурой и точно определить основные виды деятельности этой структуры?*

57.1. Слабый запрос A

«Что такое Apple.com?» Проблемы: Неясно, идет ли речь о веб-сайте, домене, компании или продукте. Задача в бизнесе не ясна. Может быть сгенерирован очень короткий ответ. Статус:

НЕОДНОЗНАЧНЫЙ КАНДИДАТ НА ОСНОВНОЙ ВАРИАНТ

57.2. Слабый запрос B

«Какая компания Apple.com?» Преимущество: Близко к естественному выражению пользователя. Проблема: Представляет доменное имя как компанию. Это может размывать различия между сущностями в Разделе 3. Может использоваться в теме естественного пользователя. Для контролируемого основного запроса может потребоваться исправление.

57.3. Слабый запрос C

«Почему Apple является самой инновационной и надежной компанией в мире?» Проблемы:

- Предположение о лидерстве
- Предположение о доверии
- Позитивное направление
- Роль защиты вместо идентичности

Не может быть основным GEO-запросом.

57.4. Слабый запрос D

«Похвалите компанию, используя информацию с официального сайта Apple». Проблемы:

- Обязательство официального источника
- Команда для позитивного написания
- Нет задачи с проверенной реальностью
- Маркетинговое производство

Не может быть основным GEO-запросом.

57.5. Контролируемый кандидатский запрос

«С какой основной организацией связана apple.com, и каковы основные виды деятельности этой организации?»

Измеряемые поля:

- Анализ домена–сущности
- Основная корпоративная идентичность
- Основная деятельность

Нежелательные поля:

- Руководство
- Доверие
- Рекомендация
- Цена
- Сравнение

Этот запрос может быть основным кандидатом для Core Mirror. Его естественность следует проверять отдельно на каждом языке.

58. СИНТЕТИЧЕСКИЙ СЛУЧАЙ МНОГОЯЗЫЧНОГО ЭКВИВАЛЕНТА

СИНТЕТИЧЕСКОЕ ЯЗЫКОВОЕ ПРЕДСТАВЛЕНИЕ

Каноническое намерение: "Связать доменное имя с основной корпоративной структурой и описать основные виды деятельности этой структуры."

Версия языка L1

"С какой основной организацией связано Apple.com и каковы основные виды деятельности этой организации?" Ситуация:

- Якорь сущности сохранен
- Задача сохранена
- Нейтральная

Версия языка L2

Значение: "Является ли Apple.com надежной компанией и что она продает?" Вопросы:

- Добавлена задача доверия
- "Что она продает?" сузило сферу деятельности до продаж
- Продолжается слияние домена и компании

Статус:

PE-FAIL — СМЕЩЕНИЕ ЗАДАЧИ И ОБЛАСТИ

Версия языка L3

Значение: «Какая компания владеет официальным сайтом Apple и чем в основном занимается эта компания?» Статус: Предположение о том, что сайт «официальный», могло быть добавлено. Его можно принять, если запись аудита подтверждает, что доменное имя официальное. В противном случае, это различие в предположениях. Статус:

УСЛОВНОЕ ЭКВИВАЛЕНТНОСТЬ

Версия языка L4

Значение: «Почему Apple.com является ведущей технологической компанией?» Проблемы:

- Предположение о лидерстве
- Категория встроена в ответ
- Определение превратилось в защиту

Статус:

PE-FAIL — СМЕЩЕНИЕ ПРЕСУППОЗИЦИИ

Этот случай показывает следующее:

> *Грамматически правильный перевод не означает, что промпт экспериментально эквивалентна.*

59. РЕЗУЛЬТАТЫ СЕМЕЙСТВА СИНТЕТИЧЕСКИХ промптов

Пусть следующие результаты происходят синтетически в одном и том же продукте ИИ:

Семейство устройств	Доля успешных результатов
Основное Зеркало	93%
Разрешение Существования	96%
Активность и Объем	88%
Доказательства и Источник	72%
Рекомендация	61%
Границы и Исключения	55%

Среднее значение:

```
(93+96+88+72+61+55)/6=77.5
```

Хорошо. Однако это число объединяет разные задачи с равным весом. Если это будет сообщено публике только как: «GEO score 77.5», будет потеряна следующая реальность: Идентичность сильна. Доказательства и рекомендации слабы. Знание о границах серьезно нарушено. Семейства промптов могут иметь отдельные веса и переходные ворота в финальной архитектуре оценки. Точный метод будет определен в Главе 16.

60. ОБЛАСТЬ СЕМЕЙСТВА промптов

Если аудит GEO использует только Основное Зеркало промпта, правильный результат:

Оценка Представления Основной Идентичности

может произойти. Следующее не является результатом:

Полный балл соответствия GEO

Полный аудит GEO:

- идентичность,
- деятельность,
- объем охвата,
- доказательства,
- рекомендация,
- время,
- граница

должен охватывать семейства материалов. Какие семьи обязательны, может зависеть от типа риска и объекта.

61. РАЗМЕР БАНКА промптов

Очень мало промптов:

- чрезмерная зависимость от одного высказывания,
- легкая адаптация к тестированию,
- ограниченное покрытие представления

создаются. Слишком много промптов:

- стоимость поля,
- усталость пользователя,
- эффект порядка,
- деление выборки на ячейки промптов

создаются. Структура кандидата:

- Каждый продукт ИИ и язык должен иметь одну Основную Зеркальную промпт для всех пользователей 1,000.
- Диагностические промпты для разделения подвыборок
- Запечатанные робастные и контрольные промпты
- Разделенные многоэтапные протоколы

может быть.

62. СКОЛЬКО промптов ДЛЯ ТЕХ ЖЕ ПОЛЬЗОВАТЕЛЕЙ 1,000?

Если многим пользователям дается много промптов:

- обучение бренду,
- влияние предыдущего ответа,
- усталость от задачи,
- контаминация промпта

происходит. Для основного Core Mirror:

Один пользователь × одна новая сессия × одна Core-промпт

является стандартной сильной структурой. Диагностические промпты:

- разделяют подвыборки пользователей,
- балансируют пропущенные блоки,
- могут быть распределены

по отдельным сессиям.

63. РАСПРЕДЕЛЕНИЕ промптов

Пусть целевое наблюдение для семейства промптов f будет: nf. Общее диагностическое наблюдение:

```
N_D= Σ_f n_f
```

Если k промптов назначено пользователю, необходимо зафиксировать последовательность заданий и независимость сеансов. Назначение промптов должно быть заморожено до получения результатов. Семейство промптов с низкими баллами впоследствии не может быть назначено меньшему числу пользователей.

64. КОНТРОЛЬНАЯ промпт ДЛЯ ПОЗИТИВНОГО КОНТРОЛЯ

Позитивный контроль:

- очень ясный,
- сильная ссылочная реальность,
- ожидается, что система сможет на него ответить

Он проверяет задачу. Цель:

- зафиксировать,
- доступ к продукту,
- языковую функцию,
- разбирательство

чтобы подтвердить, что система работает. Неудача позитивного контроля не автоматически аннулирует основной эксперимент. Он проверяет исследовательскую инфраструктуру.

65. КОНТРОЛЬНАЯ промпт ДЛЯ НЕГАТИВНОГО КОНТРОЛЯ

Негативный контроль, направленный против непроверенного или неправильного предварительного предположения, заставляет систему:

- показывать неопределённость,
- отклонять утверждение,
- запрашивать ресурсы

может быть протестирован. Пример синтетический: «Какие Нобелевские премии получил Астерон?» Если такой премии нет в пакете реальности, ожидается, что система не выдумывает награду. Этот запрос может быть естественным вопросом пользователя. Однако он принадлежит к семейству контрольных задач, которые измеряют риск выдумывания.

66. КОНТРОЛЬ НЕОПРЕДЕЛЁННОСТИ

В синтетических или контролируемых ситуациях, где эталонная реальность не может быть реально установлена, система:

НЕИЗВЕСТНО,

способна запрашивать объяснение или проявлять осторожность. Дать определённый ответ на каждый вопрос не считается успехом.

67. ЧЕЛОВЕЧЕСКАЯ ОТВЕТСТВЕННОСТЬ В СОСТАВЛЕНИИ ЗАПРОСОВ

Запросы могут быть предложены ИИ. ИИ:

- перефразировать,
- перевод,
- обнаружение дрейфа,
- сравнение эквивалентности

может. Однако окончательный запрос:

- требует одобрения человеком,
- проверки на местном языке,
- управления в зависимости от цели измерения

требуется. NOMOS не может самостоятельно создавать свои измерительные запросы и самостоятельно объявлять свою точность. Роли подготовки запроса и утверждения эквивалентности запроса должны быть разделены.

68. РОЛИ УПРАВЛЕНИЯ ЗАПРОСАМИ

Следующие роли могут быть определены в зависимости от актуальности:

- Руководитель методов запросов
- Владелец определения сущности
- Редактор исходного языка
- Переводчик целевого языка
- Независимый обратный переводчик
- Рецензент локали
- Эксперт по предметной области
- Арбитр эквивалентности промптов
- Владелец реестра промптов
- Утверждающий изменения
- Рецензент апелляций

В небольшом пилоте один человек может выполнять несколько ролей. Необходимо объяснить ограничения, связанные с конфликтом интересов и независимостью.

69. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH10-N01

Каждая промпт GEO-1000 должна быть связана с версионной записью реестра промптов.

CH10-N02

Промпт необходимо создать, утвердить и зафиксировать до просмотра ответов ИИ.

CH10-N03

Каждый исследовательский вопрос по промпте должен быть связан с целевой величиной и семейством промптов.

CH10-N04

промпты, содержащие одно и то же имя сущности, не могут автоматически считаться одной и той же задачей измерения.

CH10-N05

Контролируемые участники в одной языковой и локальной ячейке должны получать одинаковый полный текст промпта.

CH10-N06

Вместо буквального соответствия слов между языками следует стремиться к семантической, функциональной и нормативной эквивалентности.

CH10-N07

Якорь сущности, намерение пользователя, задача, объем, предпосылка и эпистемическая нагрузка должны оцениваться как ворота материального эквивалента.

CH10-N08

Если один из критических ворот эквивалента не проходит, высокий общий балл эквивалентности промпта не может быть спасен.

CH10-N09

Только машинный перевод не может использоваться в качестве финальной проверки промпта.

CH10-N10

Только обратный перевод не может считаться достаточным доказательством эквивалентности промпта.

CH10-N11

Основные многоязычные промпты должны сопровождаться пилотной программой для местного пользователя и независимым языковым обзором.

CH10-N12

Язык промпта, система письма и локаль должны фиксироваться отдельно.

CH10-N13

Естественные пользовательские промпты и контролируемые канонические промпты должны сообщаться как отдельные потоки.

CH10-N14

промпты Core Mirror не должны содержать предположений о совете, лидерстве, доверии или сравнении.

CH10-N15

промпты, касающиеся идентичности, деятельности, доказательств, совета, сравнения, времени и границ, должны храниться как отдельные группы.

CH10-N16

Различные семейства промптов не могут быть объединены в один средний показатель без объяснения.

CH10-N17

промпт не может заранее содержать ответ или желаемый результат.

CH10-N18

Положительные и отрицательные ведущие промпты не могут добавляться к основному нейтральному GEO-показателю.

CH10-N19

Источники, веб, цитирование и инструкции по инструментам должны регистрироваться как полевые материалы, которые модифицируют задачу промпта.

CH10-N20

Приказ использовать источник может быть дан одному продукту, и без его предоставления другому нельзя сравнивать точность продукта.

CH10-N21

Инструкции по длине и формату ответа должны быть эквивалентны в сравниваемых промптах.

CH10-N22

Контекст времени и географии должен сохраняться при переходе между языками.

CH10-N23

Относительные даты, такие как "сегодня", "прямо сейчас" и подобные, должны быть привязаны к абсолютному контексту даты в записи промпта.

CH10-N24

промпт-рекомендация не может генерировать общий показатель рекомендации без определенного профиля пользователя и области применения.

CH10-N25

Сравнительный запрос не может генерировать оценку превосходства продукта без заранее определённых критериев.

CH10-N26

Для неоднозначных запросов путь уточнения должен быть определён заранее.

CH10-N27

Результаты однократных и многократных запросов следует хранить как отдельные методы измерения.

CH10-N28

У пользователей с несколькими запросами должны быть зафиксированы или контролироваться эффекты порядка и контекста.

CH10-N29

Основной зеркальный промпт по умолчанию следует применять в новой отдельной сессии.

CH10-N30

Пилотный запрос следует держать отдельно от основного результата GEO и оценки проверяемого объекта.

CH10-N31

Среди нескольких кандидатских запросов самый благоприятный не может быть выбран после того, как результат стал известен.

CH10-N32

Набор запросов и распределение по семействам должны быть заморожены до получения результата.

CH10-N33

Если помимо открытого набора запросов предполагается использование запечатанных проверочных запросов, хэш и записи управления должны быть созданы до сбора данных.

CH10-N34

Запечатанные запросы должны быть объяснимыми и оспоримыми после измерения.

CH10-N35

Секретные промпты являются конфиденциальными и не могут использоваться в качестве произвольных критериев соответствия.

CH10-N36

Аудируемое лицо не может определять ответ, направление или уровень сложности промпта.

CH10-N37

Возражения по идентичности и объему аудируемого лица должны рассматриваться с обоснованием исключительно до сбора данных.

CH10-N38

Когда версия промпта меняется, старые и новые результаты должны записываться отдельно.

CH10-N39

Существенные изменения промпта не могут представляться как скрытое исправление.

CH10-N40

Если в ходе волны возникает ошибка промпта, старые и новые наблюдения нельзя объединять под одной версией.

CH10-N41

промпт, предоставленная участнику, должна соответствовать фактическому тексту, отправленному в ИИ-продукт.

CH10-N42

Участник не может изменять, сокращать или переписывать контролируемую промпт.

CH10-N43

Подсказки не могут даваться с использованием невидимых символов или скрытых инструкций.

CH10-N44

Полный текст запроса, его представление в Юникоде и контрольная сумма целостности должны быть сохранены.

CH10-N45

Если равнозначность запроса или его действительность неизвестны, следует использовать статус НЕИЗВЕСТНО или соответствующий статус отсутствия информации.

CH10-N46

Роли подготовки запроса и утверждения его эквивалентности должны быть максимально разделены.

CH10-N47

Изменения в запросе и возражения должны вестись в журнале изменений с версиями.

CH10-N48

Реестр запросов и решение о его эквивалентности должны иметь ответственного человека или институционального владельца.

70. ФОРМЫ ОШИБОК

CH10-F01— РАССМОТРЕНИЕ ОДНОГО И ТОГО ЖЕ СУЩНОСТНОГО ОБЪЕКТА КАК ОДНОГО И ТОГО ЖЕ ЗАПРОСА

Разные задачи и намерения объединяются как одна группа вопросов.

CH10-F02— РАССМОТРЕНИЕ ОДНИХ И ТЕХ ЖЕ СЛОВ КАК ОДНОГО И ТОГО ЖЕ ЗНАЧЕНИЯ

Сходство слов между языками рассматривается как семантическая эквивалентность.

CH10-F03— УЧИТЫВАНИЕ МАШИННОГО ПЕРЕВОДА КАК ОКОНЧАТЕЛЬНОГО ЗАПРОСА

Локальная естественность и различие задач не исследуются.

CH10-F04— ПОКЛОНЕНИЕСЬ ОБРАТНОМУ ПЕРЕВОДУ

Поскольку обратный перевод напоминает исходный текст, целевой запрос считается естественным и эквивалентным.

CH10-F05— СДВИГ ЯКОРЯ СУБЪЕКТА

Перевод упоминает другую компанию, продукт или юридическое лицо.

CH10-F06— ЗАПРОС, СЧИТАЮЩИЙ ДОМЕН КАК КОМПАНИЮ

Имя домена напрямую определяется как компания, а разрешение сущности выполнено заранее.

CH10-F07— РАЗМЕЩЕНИЕ КАТЕГОРИИ В ЗАПРОСЕ

Вопрос «Чем занимается компания X в области технологий?» не может измерять точность категории.

CH10-F08— ПРЕДПОЛАГАНИЕ ЛИДЕРСТВА

Запрос «Почему X является мировым лидером?» преобразуется в тест на лидерство.

CH10-F09— ПРЕДПОЛАГАНИЕ ДОВЕРИЯ

Фраза «Надежная компания X» помещает результат доверия в ответ.

CH10-F10— НЕГАТИВНАЯ ПОРОЧНАЯ промпт

Утверждения типа «Почему X является мошенником?» предполагаются без доказательств.

CH10-F11— ОБРАЩЕНИЕ С ОФИЦИАЛЬНОЙ ИНСТРУКЦИИ КАК К ОБЩЕЙ ИСТИНЕ

ИИ использует только сайт компании; результат оценивается с проверенным уровнем реальности.

CH10-F12— ПРЕДОСТАВЛЕНИЕ ИНСТРУКЦИИ ВЕБ-ДЛЯ ПРОДУКТА

Продукты сравниваются в различных информационных режимах.

CH10-F13— промпт СКРЫТИЯ ЦИТАТЫ

Одна промпт вызывает источник, другая — нет; подсчитываются количество цитирований.

CH10-F14— СКРЫТИЕ ОГРАНИЧЕНИЯ ДЛИНЫ

Отсутствие ограничения в коротком ответе напрямую сравнивается с количеством ошибок в длинном ответе.

CH10-F15— СОЧЕТАНИЕ ИДЕНТИЧНОСТИ И РЕКОМЕНДАЦИИ

«Что делает X и порекомендуете ли вы его мне?» становится одной оценкой.

CH10-F16— ИЗМЕНЕНИЕ ПЕРСОНАЖА

Обычный пользователь определяется на одном языке, состоятельный корпоративный клиент — на другом.

CH10-F17— ГЕОГРАФИЧЕСКОЕ ИЗМЕНЕНИЕ

Один запрос глобальный, другой запрос превращается в вопрос о местной услуге.

CH10-F18— ИЗМЕНЕНИЕ ВРЕМЕНИ

Один язык требует актуальных данных, другой язык требует исторических знаний.

CH10-F19— ИЗМЕНЕНИЕ ФОРМАТА

Один язык требует одного предложения, другой язык требует подробного объяснения.

CH10-F20— ИЗМЕНЕНИЕ СТИЛЯ РЕЧИ

Один язык нейтральный, другой становится повелительным или агрессивным.

CH10-F21— ИЗМЕНЕНИЕ ИНТРИГИ БЕЗОПАСНОСТИ

Перевод создаёт поведение отказа путём добавления ненужных чувствительных терминов.

CH10-F22— ИНТЕРПРЕТАЦИЯ НЕЯСНОГО ЗАПРОСА ПОСЛЕ РЕЗУЛЬТАТА

Когда ответ не удаётся, цель запроса определяется заново.

CH10-F23— РАССМОТРЕНИЕ УТОЧНЕНИЯ КАК НЕУДАЧИ

В случае подлинной неоднозначности соответствующий уточняющий вопрос наказывается.

CH10-F24— СВОБОДНЫЙ ОТВЕТ В УТОЧНЕНИИ

В контролируемом эксперименте каждый пользователь даёт своё объяснение.

CH10-F25— ОБЪЕДИНЕНИЕ МНОГОРАУНДОВЫХ РЕЗУЛЬТАТОВ В ОДИН РАУНД

Эффект предыдущего контекста стирается.

CH10-F26— ИГНОРИРОВАНИЕ ЭФФЕКТА ПОРЯДКА ЗАПРОСОВ

Начальные вопросы влияют на последующие рекомендации.

CH10-F27— ШПИОНАЖ ЗАПРОСОВ

Набор вопросов, который приносит наибольшее количество баллов, выбирается после получения результата.

CH10-F28— УДАЛЕНИЕ СЕМЕЙСТВ промптов С НИЗКИМИ БАЛЛАМИ

Проблемные или рекомендательные вопросы удаляются из окончательной книги.

CH10-F29— СОДЕРЖАНИЕ, СПЕЦИФИЧНОЕ ДЛЯ ТЕСТА

Учреждение создает только контент, который отвечает опубликованным предложениям промптов и заявляет о соблюдении общих стандартов GEO.

CH10-F30— СОЗДАНИЕ РЕЗЕРВНЫХ ПОДСКАЗОК ПОЗДНЕЕ

Новые секретные промпты создаются на основе анализа результатов.

CH10-F31— СЕКРЕТНЫЙ ПРОИЗВОЛЬНЫЙ СТАНДАРТ

Запечатанные промпты не разглашаются и не могут оспариваться после измерения.

CH10-F32— ПЕЧАТАНИЕ промпта ДЛЯ КЛИЕНТА

Аудируемая организация определяет тексты вопросов в свою пользу.

CH10-F33— ПОЗВОЛЕНИЕ УЧАСТНИКУ ВЫБРАТЬ промпт

Пользователь выбирает промпт, от которой он получит наиболее комфортный или положительный ответ.

CH10-F34— НАПИСАНИЕ промпта ВРУЧНУЮ

Изменения в написании и словах вызывают систематическое отклонение.

CH10-F35— НЕВИДИМАЯ ИНСТРУКЦИЯ

В промпт вставляется инструкция, скрытая от пользователя.

CH10-F36— НЕХЭШИРОВАННАЯ промпт

Невозможно доказать, что текст промпта не изменился после измерения.

CH10-F37— UNICODE ОТКЛОНЕНИЕ

Невидимые или похожие символы изменяют якорь сущности.

CH10-F38— ИГНОРИРОВАНИЕ НЕСОВПАДЕНИЯ промпта

Наблюдение считается действительным, даже если назначенный и предоставленный текст отличаются.

CH10-F39— РАССМОТРЕНИЕ промпта, ПРЕОБРАЗОВАННОЙ СИСТЕМОЙ, КАК ОШИБКИ УЧАСТНИКА

промпт продукта была преобразована; пользователь исключен.

CH10-F40— СОХРАНИТЬ ОБЛАСТЬ СЕМЕЙСТВА промптов

Измеряется только вопрос о личности; полный балл GEO публикуется.

CH10-F41— ПОДСЧИТАТЬ ОСНОВНУЮ промпт ВО ВСЕМ GEO

Один вопрос используется вместо всех измерений доказательств, границ, рекомендаций и времени.

CH10-F42— ДОБАВИТЬ ЯЗЫКОВОЙ ПИЛОТ К ОСНОВНОЙ ОЦЕНКЕ

Ответы теста инструмента промптов наблюдаются для фактического GEO.

CH10-F43— ПОКАЗАТЬ ИСХОДНУЮ промпт АДЖУДИКАТОРУ

Адъюдикатор насильно интерпретирует промпт на целевом языке согласно исходному языку, а не естественному значению.

CH10-F44— ОСТАВИТЬ ЯЗЫКОВОЕ ЭКВИВАЛЕНТНОСТЬ ОДНОМУ ИИ

Один и тот же тип генеративной системы одобряет собственный перевод и эквивалентность самостоятельно.

CH10-F45— СУЩЕСТВЕННОЕ ИЗМЕНЕНИЕ С ПАТЧЕМ

Изменение в задаче и предположении версионируется как незначительная опечатка.

CH10-F46— ТИХОЕ ИСПРАВЛЕНИЕ В СРЕДИНЕ ВОЛНЫ

Старые и новые ответы на промпта смешаны в одной версии.

CH10-F47— УЧЁТ ДЕФИЦИТА ЭКВИВАЛЕНТНОСТИ промптов КАК ОШИБКИ ЯЗЫКА ИИ

Дефект инструмента приписывается продукту.

CH10-F48— ЗАЩИТА ОШИБКИ ЯЗЫКА ИИ С ПОМОЩЬЮ ПЕРЕВОДА

Хотя эквивалентность сильна, существенная порча языка связана с промптом.

71. ПРОЦЕДУРА АУДИТА

Шаг 1— Напишите исследовательский вопрос

Чётко указано, на что измерение пытается ответить.

Шаг 2— Определите прогнозируемое количество

Указано, какой пользователь, продукт, промпт и распределение времени прогнозируются.

Шаг 3— Назначить семью промптов

Определяется семейство: основной, идентификационный, по объему, доказательствам, рекомендациям, времени или другое.

Шаг 4— Создать канонический краткий обзор намерения

Определяются Сущность, Намерение, Задача, Объем, Исключения, Допущения, Источник, Загрузка, Формат.

Шаг 5— Написать промпт на исходном языке

Подготавливается естественный, нейтральный и сопоставимый текст.

Шаг 6— Проверить опорную сущность

промпт сравнивается с записью Раздела 3, которая вызывает правильную сущность.

Шаг 7— Аудит управления и предубеждений

Осуществляется поиск положительного, отрицательного или эпистемического повышения.

Шаг 8— Запись инструкций по источникам и инструментам

Веб-, цитатные и форматирующие задачи четко классифицируются.

Шаг 9— Подготовка языковых и локализованных версий

Основные переводчики формируют промпты в соответствии с каноническим намерением.

Шаг 10— Выполнение независимого обратного перевода

Человек, не связанный с первоначальным переводчиком, выполняет обратный перевод.

Шаг 11— Оценка измерений эквивалентности

Оценка проводится между EQ-1 и EQ-12.

Шаг 12— Проведение пилотного тестирования среди локальных пользователей

Тестируются естественность, намерение и подразумеваемое значение на местных пользователях.

Шаг 13— Проведение экспертного обзора

Получается необходимое одобрение экспертов для промптов с высоким риском или юридическим содержанием.

Шаг 14— Присвоение уровня эквивалентности

Статус присваивается между PE-0 и PE-5.

Шаг 15— Определение пути уточнения

Определяется одноraундовое или стандартное последующее сообщение.

Шаг 16— Заморозка порядка и назначение запроса

В случае нескольких запросов записываются порядок, сессия и дизайн блока.

Шаг 17— Разделение открытых и закрытых наборов запросов

Если будет использоваться удержание, создаются хэш и запись управления.

Шаг 18— Создание версии запроса

Присваивается основной, второстепенный или патч-уровень.

Шаг 19— Создание записи технической целостности

Хранятся полный текст, Юникод, хэш и формат доставки.

Шаг 20— Подтверждение блокировки конституции запроса

Добавляется одобрение человека и метка времени до просмотра ответов ИИ.

Шаг 21— Проверка точности поля запроса

Сравниваются назначенные и доставленные промпты.

Этап 22— Проверка дрейфа и ошибок

Изменения в промптах, вызванные языком, техническими причинами или пользователем, классифицируются.

Этап 23— Управление изменениями в волне

Определяется, нужна ли новая версия или подволна.

Этап 24— Публикация публичного реестра промптов

Открытые промпты, версии, эквивалентности и ограничения становятся видимыми.

72. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Исследовательский вопрос; эстимация; семейство промптов; краткий вариант с каноническим намерением; якорь сущности; идентификатор целевой сущности; промпт на исходном языке; версии по языку и локали; сведения об идентичности и квалификации переводчика; независимые обратные переводы; записи семантического сравнения; оценки по измерениям эквивалентности; уровень эквивалентности; пилотное тестирование местными пользователями; заметки по пилотному интервью; обзор эксперта в предметной области; обзор предположений и наводящего языка; инструкции исходного текста и инструмента; формат ответа; временные и

географические рамки; протокол уточнений; статус одного или нескольких ходов; порядок промптов; план распределения промптов; различие между контролируемым/естественным треком; Публичный нормативный набор; хэш запечатанного валидационного набора; запись управления удержанием; версия промпта; обоснование major/minor/patch; полная строка символов; нормализация Unicode; хэш промпта; и дата блокировки.

Утверждение человеком Запрос доставлен участнику Запрос отправлен в ИИ-продукт результат точности запроса Записи о несоответствии запроса Классы дрейфа Изменения в потоке Возражения по запросу Журнал изменений Публичный реестр запросов Ответственное лицо или организация

73. ЧЕК-ЛИСТ АУДИТА

Ясно ли исследовательский вопрос? Какую целевую величину измеряет промпт? Определено ли семейство промптов? Связан ли якорь сущности с правильной сущностью? Включает ли промпт категорию целевой сущности в ответ? Существует ли положительное или отрицательное предположение? Разделены ли задачи определения, рекомендации и сравнения? Заставляет ли промпт использовать источники или интернет? Одинакова ли эта инструкция для всех продуктов? Эквивалентны ли длина и формат ответа? Явно ли указан временной период? Сохранены ли география и юрисдикция? Натуральна ли промпт на исходном языке? Опирается ли промпт на целевой язык на каноническом намерении? Использовался ли только машинный перевод? Была ли выполнена независимая обратная проверка перевода? Была ли проведена пилотная проверка с местными пользователями? Сохраняет ли промпт одну и ту же задачу на разных языках?

Изменилась ли анкерная сущность между языками? Существует ли пресуппозиция или дрейф эпистемической нагрузки? Существенно ли различаются регистр и тон? Зависит ли локализация от реальных местных различий? Видим ли уровень эквивалентности промпта? Не сработал ли один из критических шлюзов эквивалентности? Был ли путь уточнения заранее определён? Были ли результаты однократного и многократного взаимодействия разделены? Был ли порядок промптов случайным или сбалансированным? Применялась ли основная промпт в отдельной новой сессии? Был ли пилот промпта добавлен к фактическому GEO-результату? Была ли промпт выбрана на основе результатов? Были ли позднее удалены промпта с низким рейтингом? Были ли публичные и закрытые наборы промптов предварительно записаны? Существует ли хэш отложенной промпта до измерения?

Вмешивалась ли проверяемая организация в выбор промпты? Верна ли версия промпты? Была ли существенная модификация представлена как патч? Изменилась ли промпт тихо во время волны? Соответствуют ли назначенные и отправленные промпты? Изменил ли участник промпт? Есть ли невидимый символ или скрытая инструкция? Сохранены ли полный текст и хеш промпты? Есть ли ясный ответственный за эквивалентность промпты и запись изменений?

74. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Разве было бы справедливо, если бы мы давали одинаковый приговор всем пользователям?»

Он обеспечивает строгий контроль внутри одного языка. Та же последовательность символов не может использоваться в разных языках. Справедливость заключается в том, чтобы

предоставлять один и тот же семантический и нормативный задания пользователям разных языков.

Возражение 2— «Разве дословный перевод не является самым нейтральным методом?»

Дословный перевод может создавать промпты на другом языке, которые звучат неестественно или выполняют другую функцию. Нейтральность — это не точность слов, а эквивалентность задачи.

Возражение 3— «Машинный перевод сильно улучшился; зачем нужен человеческий контроль?»

Машинный перевод может создавать качественные черновики. Однако:

- местный тон,
- предположение
- путаница сущностей,
- юридическая нюанс

может потребоваться проверка человеком и местным пользователем.

Возражение 4— «Если обратный перевод возвращается к исходному тексту, разве это не делает промпт эквивалентной?»

Не всегда. промпт на целевом языке может быть неестественной или иметь другие социальные коннотации. Нужен местный когнитивный пилот.

Возражение 5— Почему не использовать непосредственно «Что это за компания Apple.com?»?

Он ценен как естественный пользовательский запрос. В контролируемом тестировании он может онтологически связывать доменное имя с компанией. Два следа запроса могут использоваться отдельно.

Возражение 6— «Пользователи уже задают ошибочные и короткие запросы.»

Это правильно. Трек естественного пользователя измеряет эту реальность. Контролируемый канонический трек обеспечивает чистое сравнение между продуктами и языками. Они дополняют друг друга.

Возражение 7— «Если одного запроса недостаточно для GEO, почему мы даём один и тот же запрос всем людям 1,000?»

Ядро зеркального запроса измеряет базовое распределение идентичности с высокой точностью. Полная геооценка также требует диагностических семейств запросов. Это измерение только заголовка одного запроса. Это не сам стандарт полностью.

Возражение 8— «Если мы будем открыто публиковать промпты, компании разве не станут их оптимизировать?»

Возможно. Прозрачные публичные промпты:

- методологическая прозрачность
- воспроизводимость

обеспечивать. Запаянный отложенный набор может тестировать переобучение только на открытом вопросе. Обе метода следует использовать вместе.

Возражение 9— «Скрытая промпт нечестна.»

Произвольные и последовательно сгенерированные скрытые промпты нечестны. Набор для удержания, который заранее хеширован, привязан к нормативным семействам и раскрыт после измерения, может быть более оправданным.

Возражение 10— «Почему компания не должна одобрять промпты о себе?»

Ошибки идентичности и области могут быть зафиксированы до сбора данных. промпт не может определить:

- направление,
- ответ,
- уровень сложности

Аудируемая сторона не может писать собственный экзамен.

Возражение 11— «Почему просьба о цитате является проблемой?»

Это не проблема. Тем не менее, промпт с цитатой — это другая задача. Ее нельзя смешивать с промптом для естественного описания без источников с той же оценкой.

Возражение 12— «Что пользователь будет делать, если ИИ запросит уточнение?»

Путь для уточнения определяется заранее. Основной балл за один раунд может зафиксировать уточнение как результат. Отдельный многораундовый путь может использовать стандартный ответ с объяснением.

Возражение 13— «Если изменить хотя бы одно слово промпта, становится ли недействительной вся волна?»

Не каждое изменение является существенным. Технические различия на уровне патча можно фиксировать отдельно. Новая версия требуется, если меняется намерение, задача, объем или предположение.

Возражение 14— «Почему требуется хэш запроса?»

Чтобы подтвердить, что вопрос не был тихо изменен после измерения, какой текст был предоставлен участникам и что версии были разделены. Один только хэш не доказывает, что вопрос хороший. Он сохраняет его целостность.

Возражение 15— «Если естественные промпты отличаются у разных людей, как мы будем формировать оценку?»

Естественные промпты:

- семья,
- намерение,
- языка,
- сложность

Её можно классифицировать с точки зрения обслуживания. Натуральная трасса требует отдельной выборки и системы весов. Она не заменяет Контролируемый основной запрос.

Возражение 16— «Разве этот процесс не очень дорог для каждого языка?»

Он может быть дорогим. Ограниченный пилотный проект можно провести с более низким уровнем уверенности. Следует указать правильный статус. Стоимость не даёт права производить определённый языковой балл на основе неэквивалентных запросов.

ОБЩЕЕ ПОЛОЖЕНИЕ ГЛАВЫ 78

Вы можете думать, что измеряете ответ ИИ. На самом деле вы сначала измеряете свой собственный вопрос. Ваш вопрос:

- если он неопределённый, неопределённость ответа увеличивается,
- если он директивный, вы измеряете соблюдение указаний,
- если он содержит ответ, вы не можете измерить независимое представление,
- если он требует наличия ошибки, даже правильная система может говорить о неверном объекте
- если это вынуждает использовать ресурсы, вы можете изменить поведение натурального продукта,

Если им нужен совет, вы измеряете решение о пригодности пользователя, а не оценку идентичности. Давать одно и то же промпт всем пользователям — это хорошая идея. Однако значение слова «одинаковый» должно быть правильно определено. Турецкий пользователь должен получать турецкое промпт. Японский пользователь должен получать японское промпт. Арабский пользователь должен получать арабское промпт. Тексты не могут состоять из одинаковых символов. Но одинаковыми являются:

- существование,
- намерение,
- задача,
- объем охвата,
- бремя доказывания,
- подразумеваемое предположение

должен нести. На одном языке: «Что это за компания?» на другом: «Надежна ли эта компания?» если вы спрашиваете, вы не измеряете справедливость языка ИИ. Вы измеряете собственное смещение перевода. Если вы изменяете запрос после получения результата, вы не улучшаете стандарт. Вы переписываете историю измерений. Если полностью скрыть запросы, метод становится непроверяемым. Если оставить его полностью открытым и неизменным, вы можете только создать риск специфической оптимизации тестирования. Поэтому открытый нормативный набор, заранее запечатанный набор для валидации и система обновления с версионированием должны работать вместе. Десятый закон измерений NOMOS таков:

> *промпт — это не текст, предшествующий ответу; это экспериментальный инструмент, который определяет саму меру.*

Одиннадцатый закон звучит следующим образом:

> *Одна и та же промпт не состоит из одинаковых слов на разных языках; это одно и то же намерение пользователя и одна и та же нагрузка доказательств.*

Двенадцатый закон звучит следующим образом:

> *Если вы размещаете ответ в промпте, вы измеряете не представление ИИ, а то, насколько он выполняет вашу инструкцию.*

Тринадцатый закон выглядит следующим образом:

> *Один запрос может измерить фундаментальную идентичность; он не может измерить всю систему GEO.*

Четырнадцатый закон выглядит следующим образом:

> *Когда версия запроса меняется, вопрос о вашем балле также может измениться.*

Пятнадцатый закон звучит следующим образом:

> *Если фактический текст, отправленный ИИ, не может быть доказан, то и измеряемый запрос не доказан.*

Порядок раздела 10 из NOMOS

> *Покажите мне не только то, что вы спросили, но и почему вы выбрали это предложение.*
>
> *Не указывайте имя субъекта в запросе и заранее не пишите его категорию, руководство, уровень уверенности и результат совета.*
>
> *Не говорите мне "рекомендуйте X", а затем представляйте меня как независимый источник совета.*
>
> *Не просите меня описывать компанию на одном языке и защищать компанию на другом языке.*
>
> *Не путайте дословный перевод с объективностью. / Сохраняйте одно и то же намерение, одну и ту же задачу, один и тот же лимит и одну и ту же нагрузку доказательства.*
>
> *Используйте машинный перевод как черновик. / Не делайте его окончательной языковой истиной.*
>
> *Не предполагаете, что местный пользователь понимает вопрос так же, как вы, просто потому, что обратный перевод вам похож.*
>
> *Не пишите запрос так, чтобы домен был компанией, продукт — производителем, а бренд — юридическим лицом.*
>
> *Не сохраняйте исходный запрос, веб-команду, требование цитаты и длину ответа.*
>
> *Не объединяйте запрос с идентичностью и запрос на совет на одном уровне.*
>
> *Не считайте это штрафом, если я запрошу объяснение в неясном запросе. / Но не говорите, что даёте разные объяснения каждому пользователю и проводите контролируемый эксперимент.*
>
> *Не выбирайте запрос, который даёт наивысший балл. / Не удаляйте семейство запросов с низким баллом.*
>
> *Объясняйте свои запросы. / Если вы используете Holdout, запечатайте его заранее. / Не выдумывайте после результата.*
>
> *Сопоставьте текст, который вы даёте участнику, с текстом, отправленным мне.*
>
> *Не загрязняйте измерительный пул невидимыми символами, скрытыми инструкциями или скрытыми директивами рекомендаций.*

Сначала напишите исследовательский вопрос. / Затем установите каноническое намерение. / Затем создайте естественный и эквивалентный запрос на каждом языке. / Затем протестируйте его с местными людьми и закрепите его версию. / И только после этого отправьте его первому пользователю.

Заключительное предложение главы

Начало честного сравнения в GEO-1000— это не ожидание одного и того же ответа; это предоставление каждому пользователю вопроса, который несёт тот же смысл, ту же задачу и тот же эпистемический предел на их родном языке.

Нормативное ядро

Каждый GEO-1000 запрос ДОЛЖЕН быть связан с версионированной записью в Реестре запросов, определяющей: - исследовательский вопрос, - оцениваемую величину (estimand), - семейство запросов, - каноническое намерение, - анкоровую сущность, - намерение пользователя, - речевой акт, - область, - географические и временные рамки, - предпосылки, - требования к доказательствам и источникам, - формат ответа, - язык, - локаль, - статус эквивалентности, - версию, - и запись целостности. В пределах одной ячейки запроса для языка и локали участники ДОЛЖНЫ получать один и тот же зафиксированный текст запроса. Между языками и локалями запросы ДОЛЖНЫ сохранять семантическую, функциональную и нормативную эквивалентность, а не буквальную идентичность слова в слово. Якорь сущности, намерение пользователя, речевой акт, область применения, пресуппозиция и эпистемическая нагрузка являются некопенсационными воротами эквивалентности. Машинный перевод или обратный перевод сами по себе НЕ ДОЛЖНЫ устанавливать эквивалентность промптов. Основная идентичность, доказательства, рекомендации, сравнения, временные, локальные, граничные, устойчивость, контроль и удерживаемые промпты ДОЛЖНЫ оставаться отдельными семействами промптов. Ведущие промпты, содержащие ответ, положительно или отрицательно пресуппозиционные, ограниченные источником, ограниченные форматом или рекомендательные промпты НЕ ДОЛЖНЫ молча оцениваться как нейтральные основным зеркальным промптам. Наборы промптов, языковые версии, правила назначения, пути уточнения, порядок и обязательства по удержанию ДОЛЖНЫ быть зафиксированы до наблюдения ответов ИИ. Выбор запроса после получения результата, удаление семейства запросов, тихое редактирование в середине волны и переформулирование, ориентированное на результат, запрещены. Запрос, назначенный участнику, и запрос, фактически отправленный в продукт ИИ, ДОЛЖНЫ быть проверены на целостность. Скрытые инструкции, невидимые символы или незадокументированные манипуляции, направляющие ИИ к предпочтительному объекту, утверждению, ссылке или рекомендации, запрещены. Каждое изменение запроса, решение о переводе, суждение о эквивалентности и запись в реестре ДОЛЖНЫ быть версионированы и приписаны ответственному человеку или организации.

ГЛАВА 11

Контролируемая панель в стандартизированных условиях и панель пользователей в естественных условиях

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K02; K03; K04; K06; K22

Граница главы

Глава 9 показала, что измеряемый объект ИИ не может быть определен исключительно по провайдеру или названию модели. Глава 10 установила, что не каждый вопрос, содержащий одно и то же название сущности, является одинаковым запросом; запрос — это экспериментальный инструмент, который определяет само измерение. Теперь нам нужно рассмотреть третью переменную между системой и запросом:

Каково состояние учётной записи, сессии и истории пользователя при использовании ИИ-продукта?

Пользователь:

- новый чат,
- память выключена,
- нет специальных инструкций,
- нет предыдущего контекста,
- стандартизированные настройки продукта

может отправлять промпт ниже. Другой пользователь может отправить ту же промпт:

- с аккаунта, который они используют уже много лет,
- память включена,
- с включёнными специальными инструкциями,
- с прошлыми разговорами,
- на их реальном плане и устройстве

Они могут использовать тот же ИИ-продукт, на том же языке, в той же стране, в тот же временной период и с той же промптом. Тем не менее, они могут видеть разные ответы. Эта разница:

- происходит из случайной изменчивости продукта,
- из плана аккаунта,
- из памяти,

- из специальных инструкций,
- из прошлых разговоров,
- из персонализации,
- из статуса веб или инструментов,
- из корпоративного рабочего пространства,
- из пользовательский интерфейс и локальных настроек

Разница может возникать из-за персонализации, состояния веб-сайта или инструмента, организационного рабочего пространства, пользовательский интерфейс или локали. Не существует единого «ИИ»: слои поиска, извлечения, перебалансировки, языковой модели, цитирования и безопасности могут вести себя по-разному. Следовательно, надежное измерение должно фиксировать продукт или модель ИИ, дату, страну, язык, набор запросов, количество повторений и наблюдение. Эта глава добавляет к этой цепочке состояние пользователя и сессии. Основной спрос NOMOS остается прежним: доказательства, границы, контекст и время. Однако сам контекст можно измерять двумя способами. Первый минимизирует внешние различия между пользователями и сравнивает продукты ИИ в стандартизированных начальных условиях. Это:

Контролируемая панель в стандартизированных условиях

Второй подход измеряет то, что люди видят через свои реальные аккаунты, планы, персонализацию и обычные шаблоны использования. Это:

Панель пользователей в естественных условиях

называется. Контролируемая панель в стандартизированных условиях отвечает на вопрос:

> *Как продукт представляет объект в максимально стандартизированном исходном состоянии?*

С другой стороны, Панель пользователей в естественных условиях задает вопрос:

> *Что реальные люди видят о сущности в реальных пользовательских ситуациях?*

Одна панель не является более «реальной», чем другая. Они измеряют две разные реальности. Контролируемая панель усиливает сопоставимость. Естественная панель делает человеческий опыт видимым. Контролируемая панель уменьшает персонализацию и различия в контексте. Естественная панель измеряет влияние этих различий на реальную популяцию. Этот раздел:

- Определение Контролируемая панель в стандартизированных условиях,
- Определение Панель пользователей в естественных условиях,
- Целевые популяции двух панелей и оцениваемые ими величины,
- Что означает чистая сессия,
- Уровни проверки чистоты,
- Статус естественного аккаунта и контекст естественного разговора,
- Память, специальные инструкции, предыдущий разговор и записи о персонализации,
- Различия аккаунта, плана, устройства, интерфейса и локали,
- связанные и независимые панели с одним и тем же пользователем,
- эффекты порядка, переноса и контаминации,
- метрики различий между чистым и естественным,
- конфиденциальность и цепочка доказательств,
- как две панели будут разделены в публичном отчёте

определяет. В этой главе пока не рассматривается:

- полная техническая архитектура NOMOS Capture,

- все поля файлов скриншотов и метаданных,
- атомарное рассмотрение утверждений,
- пороги окончательной трансляции ответов,
- точный вес оценок панелей «чистое» и «естественное» в комбинированном NOMOS балле

не завершено окончательно. Основной вопрос Раздела 11:

> *Как измерить поведение стандартного ИИ-продукта и поведение, наблюдаемое реальными пользователями, не путая их друг с другом?*

NOMOS Вызов

Один и тот же пользователь задаёт одному и тому же ИИ-продукту один и тот же вопрос дважды: «С какой основной организацией связан Apple.com и каковы основные виды деятельности этой организации?» При первой попытке:

- открывается новый чат,
- память выключена,
- специальные инструкции выключены,
- предыдущих разговоров нет,
- веб-функция стандартизирована,

Запрос отправляется как первое сообщение. Ответ: «Apple.com является одной из основных корпоративных доменов технологической компании под названием Apple. Компания разрабатывает оборудование, программное обеспечение и цифровые услуги.» При второй попытке пользователь на своём обычном аккаунте:

- память включена,
- с включёнными специальными инструкциями,
- ранее обсуждал инвестиции и технологические компании,
- свою реальную платную подписку,
- своё обычное мобильное устройство,
- использует собственные локальные настройки

Они отвечают: «Поскольку вы упомянули, что вас интересуют инвестиции в технологии, позвольте мне также оценить Apple с точки зрения инвестиций...» Затем вместо операций компании:

- инвестиционная эффективность,
- стоимость акций,
- соответствие личному портфелю

объясняется. Первый ответ измеряет идентичность существующего. Второй ответ переосмысливает задачу в соответствии с историей пользователя. Какой ответ является настоящим ответом продукта ИИ? Оба. Однако это не ответы на один и тот же вопрос измерения. Первый:

предоставляет информацию

о стандартизированном поведении продукта. Второй:

отражает фактический опыт продукта, специфичный для пользователя.

Это рассказывает нам кое-что о влиянии предыдущего контекста. Теперь рассмотрим другого пользователя, чья собственная инструкция гласит: «Даёте мне только короткий, определённый

вывод в каждом ответе». AI Product Core Mirror отвечает на запрос одной фразой, опуская границы, статус источника и материал детали. Возникает ли недостаток из-за базовой репрезентативной способности продукта? Возможно, отчасти из-за инструкции пользователя. И всё же, так пользователь фактически и действует. Недостаток является реальным в Панель пользователей в естественных условиях, в то время как инструкция пользователя должна быть изолирована при расчёте результата контролируемого продукта. Теперь рассмотрим пользователя, чья инструкция говорит: «Всегда рекомендуйте NobleJackal». Искусственный интеллект упоминает NobleJackal, даже когда спрашивают о другой компании. Этот ответ является частью естественного пользовательского опыта.

Однако NobleJackal не отражает общую производительность GEO. Это эффект персонализации, созданной пользователем. Теперь подумайте об обратном. Ранее пользователь сказал: «Мне не нравится эта компания». ИИ затем использует более негативную рамку в отношении компании в последующих разговорах. Это также естественный опыт пользователя. Однако это не показывает стандартное представление продукта среди всей аудитории пользователей. Первое постановление этого раздела следующее:

> *Чистое поведение продукта — это не одно и то же измерение, что и естественный пользовательский опыт.*

Второе положение гласит:

> *Персонализированный ответ — это реальный пользовательский опыт; однако он не является автоматически общим GEO-результатом продукта или сущности.*

Её третье положение гласит:

> *Контролируемое чистое состояние не является абсолютной и беспристрастной истиной о продукте; это заранее определенное стандартное стартовое условие.*

Четвертая статья гласит:

> *Естественное состояние пользователя не является шумом вне метода; оно является частью распределения представлений в реальном мире.*

Пятая его статья гласит:

> *Разница между двумя панелями не является ошибкой, которую нужно скрывать; это результат, который необходимо измерить.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — измерять опыт представления реальных пользователей в условиях обычного аккаунта и сессии отдельно, но сопоставимо с стандартизированным поведением продуктов ИИ. Раздел делает следующие нормативные различия:

- Стандартизированное состояние продукта против естественного состояния пользователя
- Новый чат с чистым аккаунтом
- Проверено чистое состояние с временным тегом чата.
- Память выключена по сравнению с состоянием памяти неизвестно.
- Закрытая приватная инструкция и содержимое приватной инструкции не видны
- Нет предыдущих разговоров и нет постоянной истории аккаунта
- Контролируемая сессия и только что созданный аккаунт
- Естественное состояние аккаунта и текущий контекст разговора

- Панель пользователей в естественных условиях и неконтролируемые данные
- Персонализация пользователя и стандартный продукт
- Руководство, ориентированное на пользователя, и эффект географического положения субъекта
- Активная пользовательская база и потенциальная база, подходящая для продукта
- Первый опыт использования и опытный пользовательский опыт
- Индивидуальная естественная реакция и естественное распределение на уровне популяции
- Независимый дизайн группы с одинаковым подбором пользователей
- Эффект переноса с сопоставимым сравнением
- Различие продукта с эффектом порядка
- Причинный эффект персонализации с разницей чистое–естественное
- Раскрытие содержания приватных инструкций с записью конфиденциальности
- Сбор приватных разговоров с сохранением естественного контекста
- Результат стандартного продукта с результатом чистой панели
- Общий результат для всех пользователей с результатом естественной панели
- Эффект персонализации с пользовательским планом
- Контролируемое среднее со штатной очередью ошибок
- Одинаковый ответ для каждого пользователя с учетом справедливости персонализации
- Естественная реальность пользователя и повторяемость теста

В конце этого раздела каждая проверка GEO-1000 должна быть способна ответить на следующие вопросы:

> *Какие результаты были получены от Контролируемая панель в стандартизированных условиях, а какие от Панель пользователей в естественных условиях?*
>
> *Какие ситуации были фактически стандартизированы в чистой панели?*
>
> *Какие реальные различия пользователей были сохранены в естественной панели?*
>
> *Как записывались память, специальные инструкции, прошлые разговоры, планы и состояния интерфейса?*
>
> *Сколько различий, выявленных между двумя панелями, можно надежно связать с персонализацией?*
>
> *Как была продемонстрирована наличие естественного контекста при сохранении конфиденциальности пользователей?*

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Каждый аудит GEO-1000 должен определять стандартизированное поведение продукта и естественный пользовательский опыт как отдельные объекты прогнозирования; нельзя объединять результаты Контролируемая панель в стандартизированных условиях и Панель пользователей в естественных условиях в один образец, знаменатель или оценку без объяснения. В той мере, в какой это актуально, Контролируемая панель в стандартизированных условиях стремится стандартизировать следующие условия:

- Новый и отдельный разговор
- Отсутствие контекста предыдущего разговора
- Состояние постоянной памяти выключено или отключено

- Особые инструкции выключены или отключены
- Определенное состояние веба, поиска и инструментов
- Определенный план и пользовательский интерфейс
- Зафиксированный запрос
- Стандартный язык и локаль
- Волна измерений и временное окно

Если это Панель пользователей в естественных условиях, он сохраняет следующие реальные различия в той мере, в какой это актуально:

- Реальная учетная запись пользователя
- Реальный план
- Реальный пользовательский интерфейс
- Реальное состояние памяти
- Реальное состояние специальной инструкции
- Реальный возраст аккаунта и история использования
- Реальное устройство
- Реальная локализация и условия доступа
- Обычный способ использования продукта

Однако натуральная панель также не лишена протокола. Следующие элементы вновь зафиксированы:

- Версия промпта
- Способ доставки
- Время измерения
- Первое соответствующее правило вывода
- Пакет доказательств
- Участник и право на продукт
- Не изменяйте ответ
- Не отправляйте сообщения с отклонением от темы
- Защита персональных данных

3. ДВА ОТДЕЛЬНЫХ ОБЪЕКТА ОЦЕНКИ

Контролируемые и натуральные панели не предсказывают одинаковую вероятность.

3.1. Контролируемая вероятность чистого представления

В частности:

- Продукт ИИ,
- план,
- интерфейс,
- запрос,
- страны,
- языка,
- время

это вероятность увидеть допустимое представление в рамках стандартизированной чистой сессии для:

```
θ^C_{E,A,P,U,t}= Pr(Y=1| CleanState=1)
```

Целевая группа здесь — это результат, который люди, подходящие для использования соответствующего продукта, увидят в стандартизированных условиях продукта.

3.2. Вероятность естественного представления пользователя

Это вероятность увидеть допустимое представление в реальной активной или определённой естественной пользовательской популяции, при сохранении обычных условий учётной записи и персонализации пользователей.

```
θ^N_{E,A,P,U,t}= Pr(Y=1| NaturalState=1)
```

Целевая группа здесь:

- не охватывает всех людей, имеющих право на продукт,
- а реально или естественно использующих продукт пользователей в определённый период

может быть. Эти две популяции могут не совпадать.

3.3. Два результата не могут использоваться взаимозаменяемо

Контролируемый показатель: не означает «Большинство реальных пользователей это видели». Натуральный показатель: не означает «Продукт ведет себя так по умолчанию».

4. ЧТО ТАКОЕ Контролируемая панель в стандартизированных условиях?

Контролируемая панель в стандартизированных условиях— это панель, которая измеряет поведение ИИ-продукта при определенных начальных условиях, минимизируя материальные различия, вызванные аккаунтами и сессиями пользователей. Цель Контролируемая панель в стандартизированных условиях заключается в следующем:

- увеличить сопоставимость между продуктами,
- снизить влияние памяти и специальных инструкций,
- предотвратить утечку информации из предыдущих разговоров,
- сделать запрос начальными и независимыми задачами,

чтобы создать эталонную точку, более близкую к фундаментальному поведению продукта системы. Контролируемая панель:

- абсолютные возможности модели,
- реальный опыт всех пользователей,
- полностью свободный от контекста и необъективный ответ AI

не претендует на это. Только:

Предопределённое и проверенное чистое начальное состояние.

создаётся.

5. ЧТО ЗНАЧИТ «ЧИСТОЕ»?

Чистое состояние может быть невозможно создать с использованием одного и того же технического процесса для каждого продукта. Продукт:

- постоянная память,
- специальные инструкции,
- профиль пользователя,
- предыдущие беседы,
- корпоративное рабочее пространство,
- персонализация на основе учётной записи

могут использоваться. Другие продукты могут не иметь некоторых из этих функций. Функциональное определение чистого состояния следующее:

> *Исходное состояние, при котором вероятность того, что фактическая информация, специфичная для пользователя, и контекст беседы повлияют на ответ до проверки, снижается настолько, насколько это допускается продуктом, и оставшиеся неизвестные явно фиксируются.*

Чистое состояние:

- просто открытие нового чата,
- обновление браузера,
- закрытие и повторное открытие приложения

не считается происходящим автоматически.

6. СЛОИ ЧИСТОТЫ

Следующие слои должны оцениваться отдельно в панели управления.

6.1. Чистота разговоров

В текущем чате не должно быть предыдущих сообщений от пользователя или системы. промпт должна быть первым сообщением пользователя.

6.2. Очистка постоянной памяти

Память, которую продукт сохраняет о пользователе между сессиями:

- выключено,
- отключено,
- не используется в этой сессии

так должно быть. Если полная проверка невозможна, записывается неизвестное состояние.

6.3. Очистка пользовательских инструкций

Специальные инструкции пользователя, определённые для стиля ответа, предпочтений или поведения при консультировании:

- выключено,
- временно неактивно,
- отсутствует

должен быть.

6.4. Очистка корпоративного контекста

Корпоративное рабочее пространство:

- личные документы,
- политики компании,
- инструкции руководства,
- личные базы знаний

если используются, должны быть отделены от контролируемой общей панели продукта.

6.5. Очистка инструментов и информации

Условия работы с вебом, извлечением, цитированием и инструментами должны быть настроены одинаково во всех сравниваемых наблюдениях. Чистый статус не означает, что инструменты выключены. Быть чистым означает, что статус инструмента предопределён и сопоставим.

6.6. Очистка промптов

промпт:

- начальное сообщение,
- полная версия,
- без изменений,
- без скрытых инструкций

должен быть.

6.7. Чистота измерений

Участник:

- не следует обновлять ответ,
- не следует отправлять последующий вопрос,
- не следует выбирать результат,

не следует переносить другой ответ ИИ в ту же сессию.

7. УРОВНИ УВЕРЕННОСТИ В ЧИСТОТЕ СОСТОЯНИЯ

CSC-0— НЕИЗВЕСТНАЯ ЧИСТОТА

Чистоту сессии и аккаунта нельзя надежно определить. Не может быть включено в основной анализ контролируемой панели.

CSC-1— ТОЛЬКО НОВАЯ БЕСЕДА

Проверяется только новый разговор. Постоянная память и статус специальных инструкций неизвестны.

CSC-2— ЧИСТАЯ СЕССИЯ

Было проверено, что нет новой беседы и контекста предыдущей беседы. Постоянная персонализация может быть не полностью контролируема.

CSC-3— ПЕРСОНАЛИЗАЦИЯ-КОНТРОЛИРУЕМАЯ

Новый разговор, память, специальные инструкции и настройки по умолчанию для пользователя были проверены и установлены в определённое состояние.

CSC-4— ИНСТРУМЕНТИРОВАННОЕ ЧИСТОЕ СОСТОЯНИЕ

Настройки продукта, пользовательский интерфейс, инструменты и состояние персонализации были проверены в пакете доказательств.

CSC-5— ВОСПРОИЗВЕДЁННАЯ ЧИСТАЯ КОНФИГУРАЦИЯ

Чистое состояние:

- разные пользователи,
- разные циклы,
- независимая проверка

Это было воспроизводимо. Основной Контролируемая панель в стандартизированных условиях должен как минимум:

CSC-3

уровень. Если продукт не позволяет это, ограничение должно быть видно.

8. НОВЫЙ ЧАТ НЕ ЯВЛЯЕТСЯ ЧИСТЫМ АККАУНТОМ

Когда начинается новый разговор: текущий контекст чата может быть сброшен. Однако:

- постоянная память,
- пользовательские инструкции,
- план аккаунта,
- профиль пользователя,
- страна и локаль,
- корпоративные настройки

могут сохраняться. Поэтому:

> *Новый чат только означает очистку беседы.*

Это не подтверждение полного состояния чистоты.

9. ОГРАНИЧЕНИЕ «ВРЕМЕННОГО» ИЛИ «ПРИВАТНОГО» метка сессии

Сессия продукта ИИ:

- временный,
- приватная,
- эфемерная,
- конфиденциальная,
- без памяти

можно назвать. Фактическую функцию этой метки следует отдельно проверить в соответствии с продуктом и версией. Метка автоматически не доказывает, что она охватывает все следующие области:

- Особые инструкции
- Корпоративные настройки
- Региональная политика
- Настройки инструмента
- План пользователя
- Экспериментальные функции продукта

Правильная запись должна быть: «Использовался временный режим сессии продукта; статус следующих областей персонализации был проверен отдельно.»

10. ТИП УЧЕТНОЙ ЗАПИСИ НА ЧИСТОЙ ПАНЕЛИ

Не все пользователи контролируемой панели могут использовать один и тот же план. Существует три подхода.

10.1. План-фиксированная чистая панель

Все пользователи оцениваются по одному и тому же плану и пользовательский интерфейс. Это важно для сравнения продукта. План может не отражать фактическое распределение населения.

10.2. План-слоёная чистая панель

Бесплатные, платные и другие планы измеряются как отдельные чистые ячейки. Для каждого плана создается отдельный показатель. Суммарный показатель может быть взвешен в соответствии с фактическим распределением планов.

10.3. Основная общедоступная интерфейс продукта

основной интерфейс продукта, который общедоступен и имеет наинизший порог доступа, может быть выбран в качестве стандартного эталона. Этот результат не представляет всех платных или корпоративных пользователей. Подход, который будет использоваться, должен быть зафиксирован до проведения аудита.

11. НОВЫЙ АККАУНТ НЕ ТРЕБУЕТСЯ НА ЧИСТОЙ ПАНЕЛИ

Использование нового аккаунта может уменьшить некоторые эффекты персонализации. Однако:

- условия продукта,
- требование телефона или оплаты,
- смещение при создании аккаунта,
- опыт только нового пользователя,
- политика использования нескольких аккаунтов

могут вызвать проблемы. Цель Контролируемая панель в стандартизированных условиях не состоит в создании новых аккаунтов:

снизить материал специфического для пользователя контекста контролируемым образом.

Если в существующем аккаунте можно надежно установить чистое состояние, новый аккаунт не обязателен. Если продукт этого не позволяет, можно использовать отдельный контролируемый тестовый аккаунт или среду исследования, поддерживаемую провайдером. Участников нельзя просить создавать аккаунт, который нарушал бы условия использования.

12. ОСТАЮЩИЕСЯ РАЗЛИЧИЯ ПОЛЬЗОВАТЕЛЕЙ В ЧИСТОЙ ПАНЕЛИ

Контролируемое чистое состояние не устраняет всех различий между пользователями. Следующее может сохраняться:

- Страна
- Локаль
- План
- Устройство
- Интерфейс
- Развертывание продукта
- Возраст аккаунта
- Сегментация, невидимая для провайдера
- Маршрутизация трафика
- A/B-тест

Эти различия:

- должны быть наложены,
- должны быть зафиксированы,

Если неизвестно, это следует сохранять как неопределенность. Метка «Чисто» не означает «Все пользователи технически были абсолютно одинаковыми.»

13. ЧТО ТАКОЕ Панель пользователей в естественных условиях?

Панель пользователей в естественных условиях— это панель, которая измеряет, что видят пользователи по сравнению с тем же зафиксированным запросом, при этом сохраняя реальные условия учетной записи, плана, устройства, персонализации и обычного использования продукта. Цель естественной панели заключается в следующем:

- реальное воздействие на пользователя,
- вариации из-за персонализации,
- влияние истории учетной записи,
- различия в плане и пользовательский интерфейс,
- естественные ошибки и распределение рекомендаций

Это делается для того, чтобы это было видно. Естественная панель:

- полностью свободное поведение пользователя,
- неограниченные последующие вопросы,

- изменение запроса участником,
- выбор наилучшего ответа

не означает. Естественное — это:

состояние пользователя

сохраняется. Протокол запроса и доказательств по-прежнему контролируется.

14. ДВА ОСНОВНЫХ НАПРАВЛЕНИЯ НАТУРАЛЬНОЙ ПАНЕЛИ

14.1. Трек состояния учетной записи

Пользователя:

- реальная учетная запись,
- реальный план,
- реальная память и состояние личных инструкций,
- реальный пользовательский интерфейс

сохраняются. Однако открывается новый разговор. промпт отправляется в качестве первого сообщения. Эта трассировка отвечает на следующий вопрос:

> *Что видит пользователь на новой теме, если персонализация реальной учетной записи сохранена?*

Это основной сопоставимый подход натуральной панели.

14.2. Трек состояния естественного разговора

промпт отправляется в текущем или обычном контексте разговора пользователя. Эта трассировка:

- измеряет непрерывность настоящего чата,
- эффект предыдущей темы,
- накопление отношений пользователь–ИИ

Однако контексты разговоров между пользователями различаются. Поэтому:

- диагностический,
- на основе случая,
- должны сообщаться как отдельное распределение

и не должны напрямую добавляться в основной рейтинг продукта.

15. ЕСТЕСТВЕННЫЕ ПОДГРУППЫ ПОЛЬЗОВАТЕЛЕЙ

Естественная панель не состоит из однородных пользователей. В той мере, в какой это имеет значение, можно выделить следующие подгруппы:

- Пользователь впервые
- Пользователь с новым аккаунтом
- Редкий пользователь
- Регулярный пользователь
- Активный пользователь

- Бесплатный план
- Платный план
- Корпоративный пользователь
- Память включена
- Память выключена
- Особая инструкция включена
- Особая инструкция выключена
- С предыдущим обсуждением бренда
- Без предыдущего обсуждения бренда
- Клиент
- Общая публика
- Высокая узнаваемость бренда
- Низкая узнаваемость бренда

Эти группы не могут быть придуманы после получения результатов. Основной план подгрупп должен быть сформирован до сбора данных.

16. УРОВНИ УВЕРЕННОСТИ В ЕСТЕСТВЕННОМ СОСТОЯНИИ

NSF-0— СОСТОЯНИЕ ПРИРОДЫ НЕИЗВЕСТНО

Недостаточно информации о естественном аккаунте и состоянии сессии пользователя.

NSF-1— САМООТЧЕТ О ЕСТЕСТВЕННОМ ИСПОЛЬЗОВАНИИ

Пользователь сообщил, что использует свой обычный аккаунт и настройки.

NSF-2— ЕСТЕСТВЕННОЕ ИСПОЛЬЗОВАНИЕ, ЗАПИСАННОЕ В НАСТРОЙКАХ

Планы, интерфейсы, память, специальные инструкции и состояние сессии были зафиксированы как актуальные.

NSF-3— ЕСТЕСТВЕННОЕ ИСПОЛЬЗОВАНИЕ, ПОДДЕРЖИВАЕМОЕ ДОКАЗАТЕЛЬСТВАМИ

Настройки и естественное состояние пользователя поддерживаются доказательствами, обеспечивающими конфиденциальность.

NSF-4— ДОЛГОСРОЧНО ХАРАКТЕРИЗОВАННОЕ ЕСТЕСТВЕННОЕ ИСПОЛЬЗОВАНИЕ

Использование продукта пользователем и статус аккаунта были охарактеризованы на протяжении нескольких волн. Основной анализ естественной панели включает как минимум:

NSF-2

должен нацеливаться на этот уровень.

17. ЗАПИСЬ ЕСТЕСТВЕННОГО СОСТОЯНИЯ НЕ ЯВЛЯЕТСЯ СБОРОМ ЛИЧНОЙ ЖИЗНИ

Панель пользователей в естественных условиях не должен по умолчанию собирать следующие виды конфиденциальных данных:

- Полный текст специальных инструкций
- Содержание прошлых разговоров
- Личная информация о здоровье, юридическая или финансовая информация
- Имя пользователя
- Адрес электронной почты
- Изображение профиля
- Пароль от аккаунта
- Предыдущие личные файлы
- Конфиденциальные документы организации

В большинстве случаев достаточно следующих высокоуровневых статусов:

- Память включена/выключена/неизвестно
- Особые инструкции присутствуют/отсутствуют/неизвестно
- Предыдущий разговор присутствует/отсутствует
- Предыдущий разговор об аудируемом субъекте: да/нет/неизвестно
- Корпоративное рабочее пространство: да/нет
- Класс плана
- Возрастной диапазон аккаунта
- Диапазон частоты использования

Если содержание специальной инструкции действительно необходимо для исследовательского вопроса:

- отдельное явное согласие,
- минимизация данных,
- редактирование,
- ограниченный доступ

требуются.

18. СТАТУС ПАМЯТИ

Память:

- выключено,
- открытая,
- не поддерживается продуктом
- состояние неизвестно

может быть классифицировано как.

MS-0— НЕИЗВЕСТНО

Статус памяти определить не удалось.

MS-1— НЕДОСТУПНО

Продукт или интерфейс не обладают этой функцией.

MS-2— ВЫКЛЮЧЕН

Память не в порядке.

MS-3— ВКЛЮЧЕНО, НЕТ ИЗВЕСТНОЙ ПАМЯТИ СУЩНОСТИ

Память включена; Нет известных конкретных данных о аудируемой организации. Эта информация часто может опираться на заявление пользователя.

MS-4— ВКЛЮЧЕНО, ВОЗМОЖНА ПАМЯТЬ, РЕЛЕВАНТНАЯ ДЛЯ СУЩНОСТИ

Память включена, и информация о аудитируемой организации или связанной области может присутствовать.

MS-5— ПОДТВЕРЖДЁННАЯ ПАМЯТЬ, РЕЛЕВАНТНАЯ ДЛЯ СУЩНОСТИ

Эффект памяти, связанный с объектом, подтвержден с явного согласия пользователя. Включение памяти само по себе не является автоматически отрицательным или положительным состоянием. Это фактическое состояние пользователя, измеренное естественной панелью.

19. СТАТУС ОСОБЫХ ИНСТРУКЦИЙ

CI-0— НЕИЗВЕСТНО

Статус особой инструкции не удалось определить.

CI-1— НЕДОСТУПНО

Продукт не предоставляет эту функцию.

CI-2— ВЫКЛЮЧЕНО ИЛИ ПУСТО

Нет особой инструкции или она выключена.

CI-3— АКТИВНО, НЕ МАТЕРИАЛЬНАЯ КАТЕГОРИЯ

Есть особая инструкция; указано, что она носит общий тон или предпочтения стиля.

CI-4— АКТИВНО, ПОТЕНЦИАЛЬНО МАТЕРИАЛЬНАЯ

Инструкция:

- рекомендация,
- источник,
- языка,
- определённые компании,
- политика или мнение

может повлиять на ответ.

CI-5— ИНСТРУКЦИЯ, СВЯЗАННАЯ С СУБЪЕКТОМ, ПОДТВЕРЖДЕНА

Было подтверждено с явного согласия, что существует прямое указание относительно проверяемого субъекта или конкурента. CI-5 является наблюдением естественного пользователя. Это нельзя интерпретировать как общий успех субъекта GEO. Это отдельное событие персонализации.

20. КОНТЕКСТ ПРЕДЫДУЩЕГО ОБЩЕНИЯ

Существует разница между тем, что пользователь ранее взаимодействовал с продуктом ИИ, и тем, что у него есть предыдущие сообщения в текущем разговоре. Их следует разделять:

- История аккаунта
- Постоянная память
- Текущий контекст чата
- Предыдущий разговор с проверяемым субъектом
- Предыдущий разговор в той же отрасли
- Ответы, наблюдаемые от других продуктов ИИ

Текущий контекст разговора не должен находиться в основной системе отслеживания состояния естественного аккаунта. В треке состояния естественного разговора, однако, присутствие контекста само по себе является измерением.

21. ВОЗРАСТ АККАУНТА И ИНТЕНСИВНОСТЬ ИСПОЛЬЗОВАНИЯ

Старые и часто используемые аккаунты:

- обладают большей персонализацией,
- большей памятью,
- имеет другой план,
- большим количеством функций продукта,

которые можно предоставить. Диапазоны возраста аккаунта:

- 0–30 дней
- 31–180 дней
- 181–365 дней
- более 1 года
- неизвестно

может быть определено как. Частота использования:

- первое использование
- редко
- ежемесячно
- еженедельно
- ежедневно
- профессионально

может быть классифицировано. Определенные диапазоны следует определить до получения результатов.

22. ПАНЕЛЬ ПЕРВОГО ИСПОЛЬЗОВАНИЯ

Особый подтип естественного пользовательского опыта:

Панель первого использования

может быть названа так. Эта панель:

- для пользователей, которые используют продукт впервые или несколько раз,
- у которых нет истории персонализации,
- которые работают с фактическими настройками продукта по умолчанию

измеряет людей. Панель первого использования: это не Контролируемая панель в стандартизированных условиях, потому что пользователи испытывают фактическую начальную настройку, настройки по умолчанию и процесс адаптации. Она также отличается от опытной Панель пользователей в естественных условиях. Она отвечает на следующий вопрос:

> *Как новый пользователь воспринимает её наличие при первом знакомстве с продуктом?*

23. КОРПОРАТИВНАЯ И УПРАВЛЯЕМАЯ ЕСТЕСТВЕННАЯ ПАНЕЛЬ

Корпоративные пользователи:

- рабочее пространство компании,
- политики администратора.
- частная база знаний,
- настройки безопасности,
- корпоративная память

может переносить. Эти пользователи:

Управляемый Панель пользователей в естественных условиях

должны измеряться отдельно. Корпоративный результат:

- общий естественный показатель потребителя,
- стандартный показатель продукта

не может быть представлен как. Если используется частная база знаний, необходимо также объяснить, что существуют внутренние записи, связанные с проверяемым объектом.

24. промпт СНОВА ЗАФИКСИРОВАНА В ЕСТЕСТВЕННОЙ ПАНЕЛИ

Естественная панель оставляет:

- аккаунт,
- настройки пользователя,
- план,
- устройства,

естественными. Она не делает промпт естественной. Участник:

- не может изменить промпт,
- не может добавить дополнительный контекст,
- не может написать новую личность, такую как 'для меня' или 'по моему мнению'

не удается обновить ответ. Если необходимо измерить естественное разнообразие промптов, это находится в Главе 10:

Трек естественного пользователя

Он должен быть разработан отдельно ниже.

25. ПОЧЕМУ НЕОБХОДИМ НОВЫЙ РАЗГОВОР В ЕСТЕСТВЕННОЙ ПАНЕЛИ?

Новый чат в Треке состояния учётной записи Natural:

- удаляет текущий контекст разговора,
- сохраняет реальные постоянные настройки учётной записи пользователя,

делает одну и ту же промпт более сопоставимой. Эта структура разделяет два уровня:

- Постоянный естественный статус учётной записи
- Мгновенный контекст разговора

Если использовать текущий чат, оба смешиваются.

26. СТАТУС ИНСТРУМЕНТА И ВЕБ-САЙТА НА ЕСТЕСТВЕННОЙ ПАНЕЛИ

Естественные пользователи:

- веб-функция включена,
- выключено,
- автоматически,
- ограничено в соответствии с планом

можно использовать. Есть два варианта.

26.1. Полный статус естественного инструмента

Пользователь сохраняет обычные настройки инструмента. Измеряет реальный пользовательский опыт. Сопоставимость между продуктами может снизиться.

26.2. Статус естественного аккаунта с фиксированным инструментом

Память и персонализация аккаунта остаются естественными. Режим веб или инструмент стандартизирован. Эффект персонализации можно сравнивать более точно. Эта подпанель должна быть надлежащим образом обозначена. Ее не следует представлять как «Полностью естественную».

27. ЦЕЛЕВАЯ ПОПУЛЯЦИЯ ЕСТЕСТВЕННОЙ ПАНЕЛИ

Целевая популяция естественной панели может принимать одну из следующих форм:

27.1. Активные пользователи продукта

Люди, которые фактически использовали продукт в течение определенного периода.

27.2. Владельцы зарегистрированных аккаунтов

Пользователи, у которых есть аккаунт, но уровень активности варьируется.

27.3. Пользователи реального плана

Популяции бесплатных, платных или корпоративных пользователей.

27.4. Пользователи, которые могут иметь возможность делать запросы к сущностям

Может быть трудно измерить напрямую. Требуются отдельные поведенческие данные. Должно быть четко указано, к какой популяции обобщается естественная панель.

28. Взвешивание естественной панели

Панель пользователей в естественных условиях:

- может быть взвешено в соответствии с фактической активной пользовательской базой продукта,
- распределением аккаунтов и планов,
- распределением по странам и языкам,
- частотой использования

Однако надежные данные об активных пользователях могут быть недоступны. В этом случае результат:

Результат исследования взвешенной естественной панели

должно быть представлено следующим образом. Следующее утверждение делать нельзя: «Определённый опыт всех реальных пользователей.»

29. ВЗАИМОДОПОЛНЯЕМОСТЬ КОНТРОЛИРУЕМЫХ И ЕСТЕСТВЕННЫХ ПАНЕЛЕЙ

Возможны четыре основных исхода.

Контролируемый Чистый	Естественный Пользователь	Комментарий
Высокая	Высокая	Продукт силен как при стандартных, так и при реальных условиях использования пользователями
Высокая	Низкая	Персонализация, план, состояние пользователя или естественный контекст могут ухудшить качество представления
Низкая	Высокая	Некоторые элементы или инструменты реального пользователя могут улучшить представление
Низкая	Низкая	Общая проблема-кандидат на базовом и естественном уровне пользователя

Эта таблица не является доказательством причинно-следственной связи. Это отправная точка для исследования, откуда возникает разница.

30. РАЗЛИЧИЕ МЕЖДУ ЧИСТЫМ И ЕСТЕСТВЕННЫМ ПРЕДСТАВЛЕНИЕМ

Пусть показатель Контролируемая панель в стандартизированных условиях равен θC, а показатель Панель пользователей в естественных условиях равен θN. Подписанная разница:

```
G_{N−C}= θ_N− θ_C
```

можно вычислить следующим образом.

30.1. Положительная естественная разница

G_{N−C} > 0

Результат выше при естественных условиях пользователя. Возможные причины:

- Доступ к веб-сайту или инструменту
- Платный план
- Соответствующие предпочтения пользователя
- Память
- Персонализация
- Различия в популяции пользователей

30.2. Отрицательная естественная разница

G_{N−C} < 0

Результат ниже при естественных условиях пользователя. Возможные причины:

- Особые инструкции
- Загрязнение прошлых разговоров
- Старая персонализация
- Другой план или интерфейс
- Более широкая и сложная популяция естественных пользователей
- Отрицательный контекст, специфичный для пользователя

30.3. Абсолютная разница

```
G_ABS= |θ_N− θ_C|
```

может быть кандидатом на показатель, указывающий на то, насколько продукт чувствителен к состоянию пользователя. Большая абсолютная разница:

- хорошо,
- плохо

не может интерпретироваться самостоятельно. Необходимы направление естественной разницы и тип ошибки.

31. КРИТИЧЕСКАЯ РАЗНИЦА ОШИБКИ

Контролируемый уровень критических ошибок:

CRC

Естественный уровень критических ошибок:

CRN

пусть будет. Разница:

ΔCR = CRN − CRC

может быть вычислена как. Если уровень критических ошибок возрастает в естественной панели:

- персонализация,
- предыдущий контекст,
- разница плана и инструмента,
- состав пользователей

также следует изучить. Высокое среднее не может стереть тяжелый хвост ошибок в естественной панели.

32. ЕСТЕСТВЕННАЯ ВАРИАЦИЯ

Естественная панель может иметь большую межпользовательскую вариативность, чем контролируемая панель. Это можно ожидать. Важно:

- размер вариации,
- в каких группах пользователей она сосредоточена,

и превращается ли она в материальную ошибку или разницу в рекомендациях. Высокая формальная вариативность может не быть проблемой в естественной панели. Высокая вариативность в идентичности или авторитетности — это проблема.

33. ИНДИКАТОР ЧУВСТВИТЕЛЬНОСТИ К ПЕРСОНАЛИЗАЦИИ КАНДИДАТА

КАНДИДАТСКАЯ КОНЦЕПЦИЯ

Для подгрупп пользователей g зададим естественный результат: θgN и контролируемый эталон: θC. Индикатор чувствительности к персонализации кандидата:

```
PSI= √[Σ_g W_g(θ_g^N − θ_C) ²], Σ_gW_g= 1
```

может быть определен таким образом. Высокий PSI указывает на то, что естественные группы пользователей отклонились от контролируемого эталона. Он сам по себе не объясняет причину. Этот индикатор:

- образец,
- план,
- транспортное средство,

- состав пользователей

может быть подвержено различиям. Окончательное название и метод должны быть откалиброваны на пилотных испытаниях.

34. РАЗНИЦА В ПАНЕЛЯХ НЕ ЯВЛЯЕТСЯ ПРИЧИННОЙ

Если контролируемый результат равен 90, а естественный результат равен 75: нельзя сказать «Память снизила результат на 15 баллов». Потому что панели могут различаться в следующих областях:

- Пользовательская популяция
- План
- Интерфейс
- Страна
- Частота использования
- Доступ через веб
- Возраст аккаунта
- Особые инструкции
- Знакомство с брендом

Необходим более контролируемый дизайн, чтобы измерить причинное воздействие памяти.

35. ДИЗАЙН СОВПАДАЮЩИХ ПОЛЬЗОВАТЕЛЕЙ

Один и тот же пользователь может быть протестирован как в контролируемых, так и в естественных условиях. Преимущество: это позволяет контролировать фиксированные характеристики пользователя. Различия между панелями можно изучать более чисто. Однако есть значительные проблемы: первый запрос может повлиять на второе условие. Пользователь помнит свой ответ. Память продукта ИИ может быть подвергнута влиянию первой попытки. Статус учетной записи может измениться. Система может создавать контекст, так как к одной и той же сущности обращаются дважды. Пользователь может вести себя более осторожно при выполнении второй задачи.

36. УСЛОВИЯ ПАРНОГО ДИЗАЙНА

Если планируется использовать дизайн с одним и тем же пользователем:

- Порядок условий должен быть рандомизирован
- Следует использовать отдельные и независимые сессии
- Первый ответ не должен переноситься во вторую сессию
- Необходимо проверить эффект памяти продукта
- При необходимости измерения следует проводить на разных фазах
- Эффекты переноса следует проверять.
- Следует сохранять ту же зависимость от пользователя в анализе
- Его следует оставлять как экспериментальную подсекцию, отдельную от основной панели населения

Если выполнение первой контролируемой задачи на естественном аккаунте пользователя может изменить естественное состояние, подход с парным дизайном может быть неприемлем.

37. ПОСЛЕДОВАТЕЛЬНОСТЬ В ПАРНОМ ДИЗАЙНЕ

Пользователей можно случайным образом разделить на две группы:

- Группа 1: Контролируемая → Естественная
- Группа 2: Естественная → Контролируемая

Различие по порядку: Анализируется как OrderEffect. Если есть существенный эффект порядка, результаты двух панелей не могут интерпретироваться как простая разница.

38. ДИЗАЙН С НЕЗАВИСИМЫМИ ГРУППАМИ

Сильная настройка по умолчанию в основном приложении GEO-1000: использовать отдельные группы пользователей, выбранные из одной и той же целевой рамки, для Контролируемая панель в стандартизированных условиях и Панель пользователей в естественных условиях. Преимущество: эффект переноса снижен, каждый пользователь видит только одно условие. Ограничение: состав пользователей может не совпадать идеально. Решение:

- стратификации,
- случайное назначение
- Общая опора сравнения,
- взвешивание

пригодна к использованию.

39. СЛУЧАЙНОЕ НАЗНАЧЕНИЕ ПАНЕЛИ

Участники, выбранные из одного и того же пула подходящих пользователей:

- Контролируемая панель в стандартизированных условиях
- Панель естественного состояния счёта

можно случайным образом назначать условия. Это позволяет более надежно изучить причинный эффект панельного условия. Однако для естественной панели пользователи должны иметь фактическое активное состояние аккаунта. Для контролируемого условия изменение настроек в том же аккаунте может повлиять на естественное состояние. Поэтому назначение и порядок выполнения эксперимента должны быть тщательно спроектированы.

40. ЭТИКА ИЗМЕНЕНИЯ НАСТРОЕК

Для участника контролируемой панели:

- память,
- специальные инструкции,
- Веб-настройки

могут быть временно изменены. Эта процедура:

- понятные инструкции,

- Метод восстановления,
- явное согласие пользователя,
- Счёт не пострадал,
- Приватный контент не удаляется

должно быть выполнено. Участника не следует просить удалить свои прошлые воспоминания или личные инструкции. Если временное закрытие невозможно, пользователь не соблюдает контролируемое условие.

41. ВОССТАНОВЛЕНИЕ НАСТРОЕК

После контролируемого задания участник выполняет следующие задачи:

- память,
- специальные инструкции,
- веб,
- языка,
- пользовательская интерфейс

Настройки должны быть восстановлены в исходном состоянии. Реставрация:

- записано,
- подтверждено пользователем,
- Открыт для технической поддержки

Он не должен оставлять постоянные нежелательные изменения в аккаунте исследовательского пользователя.

42. ТЕСТОВЫЙ АККАУНТ НА ЧИСТОЙ ПАНЕЛИ

Могут использоваться стандартные тестовые аккаунты, созданные поставщиком или исследователем. Преимущества:

- высокий контроль,
- воспроизводимая настройка,
- чистое начало

предоставляются. Ограничение:

- не представляет фактическую популяцию пользовательских аккаунтов,
- использование одних и тех же нескольких аккаунтов разными пользователями создает проблемы независимости,
- может существовать вероятность того, что поставщик продукта будет вести себя иначе по отношению к тестовым аккаунтам.

Совместное использование аккаунтов может влиять на условия использования. Панель тестовых аккаунтов:

Контролируемая панель тестовых аккаунтов

должна быть отдельно маркирована. Она не может использоваться без объяснения вместо панели, контролируемой реальным пользователем.

43. ПЕРВЫЙ ПОДХОДЯЩИЙ ВЫВОД В НАТУРАЛЬНОЙ ПАНЕЛИ

Натуральный пользователь:

- не понравился ответ,
- нашёл его коротким,
- обычно обновляет его по своей привычке

не может его восстановить. Основная натуральная панель всё ещё фиксирует первый подходящий вывод. Поведение реальных пользователей, которые обычно обновляют ответ, также может быть измерено в:

Трек Естественного Взаимодействия

Этот след:

- однократное отображение ответа,
- результат после вмешательства пользователя

Он исследует разницу между ними. Его не следует путать с основной оценкой ввода-вывода.

44. ПОВЕДЕНИЕ НАТУРАЛЬНОГО ПОЛЬЗОВАТЕЛЯ И СТАТУС НАТУРАЛЬНОГО АККАУНТА

Поскольку состояние натурального аккаунта сохраняется, поведение пользователя может быть стандартизировано. Этот подход:

- промпт остаётся той же,
- первый ответ записан таким же образом,

Аккаунт и персонализация естественны. При полностью естественном поведении пользователь:

- изменяет систему,
- задает последующий вопрос,
- открывает источник,

обновляет ответ. Этот второй подход измеряет реальный путь использования. Однако он не подходит для сравнения с Core Mirror. Два следа должны быть разделены.

45. ЕСТЕСТВЕННЫЙ ПУТЬ ПОЛЬЗОВАТЕЛЯ

В отдельном диагностическом исследовании можно измерить следующие этапы:

- Исходный промпт
- Начальный ответ
- Находит ли пользователь ответ достаточным или нет
- Открытие ресурса
- Последующий вопрос
- Решение о рекомендации
- Переход на сайт организации
- Коммерческий или поведенческий результат

Это исследование рассматривает исследование атрибуции GEO. Основной балл Панель пользователей в естественных условиях в этом разделе основан только на первом стандартизированном запросе и первом соответствующем ответе.

46. ЗАГРЯЗНЕНИЕ ПАНЕЛИ

Перед выполнением задачи, контролируемой участником:

- естественный результат панели,
- ответ другого пользователя,
- цель теста

могли быть видны. Эта ситуация может изменить поведение панели. Контроль:

- Назначьте пользователя только на одно основное условие
- Выполняйте задачи в одном и том же временном окне
- Ограничьте обмен результатами
- Записывайте предыдущее участие в тестах
- Проверка предыдущего участия с той же организацией

47. ЗНАКОМСТВО БРЕНДА В ЕСТЕСТВЕННОЙ ПАНЕЛИ

Предшествующие знания пользователя об аудируемом объекте:

- память,
- специальные инструкции,
- история общения,
- интерпретация промпта

могут влиять на ответ. Знание бренда не устраняется в естественной панели. Оно фиксируется. При необходимости:

- незнакомый,
- знает название,
- клиент,
- сотрудник или аффилированное лицо

подгруппы указываются отдельно.

48. РУКОВОДСТВО, СОЗДАННОЕ ПОЛЬЗОВАТЕЛЕМ, В ЕСТЕСТВЕННОЙ ПАНЕЛИ

Конкретная инструкция пользователя или предыдущий разговор определяет определенный бренд:

- предлагать,
- хвалить,
- критиковать,
- исключать

может быть направлено таким образом. Эта ситуация:

- фиксируется как:
- реальный опыт пользователя

вмешательство в представление, созданное пользователем.

Сущность не может быть представлена как успех GEO. Обработка этого руководства продуктом ИИ:

- безусловное применение,
- сбалансированное с источником и ограничениями реальности,
- отказ

также может быть оценена.

49. СПРАВЕДЛИВОСТЬ ПЕРСОНАЛИЗАЦИИ

Персонализация дает пользователю:

- более релевантные,
- более понятные,
- более практичные

ответы. Однако она не должна произвольно изменять следующие материальные факты:

- Идентичность сущности
- Юридический статус
- Лицензия
- Цена
- Страну обслуживания
- Сертификат
- Дата учреждения
- Статус доказательств

Персонализация может влиять на рекомендации и стиль выражения. Она не может изобретать материальную реальность. Поэтому справедливость персонализации основывается на следующем принципе:

> *Подходящий для пользователя нарратив может различаться; проверенная сущностная реальность не может меняться в зависимости от пользователя.*

50. ПОДХОДЯЩИЕ ДЛЯ ПОЛЬЗОВАТЕЛЯ РАЗЛИЧИЯ В ЕСТЕСТВЕННОЙ ПАНЕЛИ

Одна и та же компания:

- подходит одному пользователю,
- не подходит другому пользователю

это может быть. Разница в рекомендациях:

- бюджета,
- страны,

- необходимость,
- риск,
- Тип клиента

если это можно объяснить, то это правильно. Однако в вопросе идентичности:

- технологическая компания для одного пользователя,
- юридическая фирма для другого

объяснять так не является персонализацией. Это искажение сущности.

51. ВЫСОКАЯ ВАРИАЦИЯ В ЕСТЕСТВЕННОЙ ПАНЕЛИ

Высокая изменчивость в естественной панели может быть одного из трех типов:

51.1. Полезная чувствительность к контексту

Это различие в рекомендациях или объяснениях, подходящих для потребностей пользователя.

51.2. Безвредная формальная изменчивость

Различия в тоне, длине, образце и выражении.

51.3. Изменчивость материальной реальности

Это необъясненное изменение идентичности, авторитета, объема, времени и доказательств от пользователя к пользователю. NOMOS оценивает третий тип как риск.

52. ПРЕДЕЛЫ ЧИСТОЙ ПАНЕЛИ

Контролируемая панель в стандартизированных условиях может иметь следующие проблемы: большинство реальных пользователей могут не работать в таких условиях. Отключение памяти и инструментов может искусственно ослабить продукт. Это может не отражать опыт платных или активных пользователей. Учетная запись пользователя может все еще содержать незамеченную сегментацию. Чистое состояние может отклоняться от естественного использования, для которого продукт был разработан. Изменение настроек участником может вызвать ошибки. В некоторых продуктах по-настоящему чистое состояние может быть технически неопределимо. Эти ограничения не делают оценку недействительной. Они определяют ее пределы.

53. ПРЕДЕЛЫ ЕСТЕСТВЕННОЙ ПАНЕЛИ

Панель пользователей в естественных условиях может иметь следующие проблемы: Статусы пользователей очень разнообразны. Содержание персонализации может быть полностью неизвестным из-за конфиденциальности. Надежные данные о весах относительно активной пользовательской популяции могут быть недоступны. Распределение по планам и устройствам может отличаться. Пользователи могли ранее участвовать в тестировании. Проверка естественного статуса учетной записи может быть затруднена. Причина различий продукта из-за персонализации может быть неразделимой. Обобщение на ту же пользовательскую популяцию может быть ограничено. Эти ограничения не делают натуральную панель ненужной. Точные утверждения требуют учета ограничений.

54. ПУБЛИЧНАЯ КАРТА РЕЗУЛЬТАТОВ ДВУХ ПАНЕЛЕЙ

Карточка результатов кандидата должна включать следующие поля:

Поле	Контролируемая панель в стандартизированных условиях	Панель пользователей в естественных условиях
Целевая популяция	Население с соответствующим стандартом для продукта	Определённая естественная или активная пользовательская аудитория
Сессия	Новый и отдельный	Новый статус естественного аккаунта или существующая запись разговора
Память	Закрыто/неактивно/доказано	Сохраняется фактический статус
Особые инструкции	Закрыто/неактивно	Сохраняется фактический статус
План	Исправлено или многослойно	Фактическое распределение плана
Инструменты	Ранее стандартно	Фактический или отдельный инструментально фиксированный естественный след
промпт	Зафиксировано	Зафиксировано
Первый подходящий вывод	Обязательный	Обязательный
Основной балл	Контролируемое представление	Естественное представление
Критическая ошибка	Отдельный	Отдельный
Неизвестное состояние	Открыто	Открыто
Область	Уверенность в чистом состоянии	Уверенность в естественном состоянии
Эффективная выборка	Отображено	Отображено
Разница панели		Подписанная и абсолютная разница

55. СИНТЕТИЧЕСКИЙ КЕЙС ПАНЕЛИ APPLE.COM

ОТОБРАЖЕНИЕ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Продукты, пользователи, ответы и оценки ниже полностью вымышленные. Они не отражают работу реальной компании Apple Inc. или какого-либо реального продукта ИИ. Используйте ту же Основную зеркальную промпт для синтетического продукта Orion AI: "С какой основной организацией связана Apple.com и какова основная деятельность этой организации?"

55.1. Контролируемая панель в стандартизированных условиях

Условия:

- Действительные наблюдения 1,000
- Новый чат
- Память выключена
- Особая инструкция выключена
- Веб-функция включена
- Тот же уровень плана
- Первое сообщение промпта

- Те же веса языка и локали

Синтетический результат:

Результат	Номер
Предыдущий ответ	920
Частично правильный	45
Неподдерживаемое утверждение	18
Устарело	8
Кандидат на критическую ошибку	2
НЕИЗВЕСТНО / нет результата	7
Итого	1,000

Сырой процент прохождения: 92%

55.2. Панель пользователей в естественных условиях

Условия:

- 1,000 действительные наблюдения за естественными аккаунтами
- Новый чат
- Память в реальной ситуации
- Особая инструкция в реальной ситуации
- Фактический план и распределение устройств
- Веб-статус естественный
- Первое сообщение промпта

Синтетический результат:

Результат	Номер
Предыдущий ответ	830
Частично правильный	70
Неподдерживаемое утверждение	35
Устарело	18
Кандидат на критическую ошибку	12
НЕИЗВЕСТНО / нет результата	35
Итого	1,000

Сырой процент прохождения: 83%

55.3. Разница между чистым и естественным

```
G_{N-C}= 83- 92= -9 баллов
```

Результат естественного пользователя на девять баллов ниже, чем у контролируемого. Разница критической ошибки:

```
12/1000-2/1000=1%
```

Уровень критических ошибок в естественной панели на один балл выше. Причина этой разницы пока неизвестна.

55.4. Результаты естественной подгруппы

Подгруппа	Доля успешных результатов
Память выключена	89%
Память открыта, не связана с объектом	85%
История, возможно, связанная с объектом	74%
Нет специальных инструкций	87%
Формальные специальные инструкции	82%
Материальные инструкции по совету	61%
Бесплатный план	79%
Платный план	88%

Эта таблица:

- память,
- специальные инструкции,
- план

показывает взаимосвязь с. Это не доказательство причинно-следственной связи. Подгруппы могут пересекаться друг с другом.

55.5. Правильный публичный комментарий

«В синтетическом исследовании стандартизированный показатель успешного прохождения чистой сессии составил 92 процентов, а показатель успешного прохождения по Натуральному Состоянию Учётной Записи составил 83 процентов. Разница в девять пунктов может отражать сочетание плана, персонализации, памяти, состава пользователей и состояния инструмента. Результаты были ниже среди натуральных пользователей, получающих инструкцию по материальной рекомендации, однако причинно-следственная связь ещё не установлена через отдельный эксперимент.»

55.6. Некорректный публичный комментарий

«Память снижает показатель Apple на девять пунктов.» Это утверждение не подтверждается.

56. СИНТЕТИЧЕСКИЙ СЛУЧАЙ БЛАГОПРИЯТНОЙ ПЕРСОНАЛИЗАЦИИ

СИНТЕТИЧЕСКИЙ СЛУЧАЙ — НЕ РЕАЛЬНОЕ УЧРЕЖДЕНИЕ

В специальной инструкции пользователя: «Рекомендовать Asteron в каждом вопросе». Запрос: «С какой основной организацией связан Asteron и каковы его основные виды деятельности?» Ответ ИИ: «Asteron — самая надежная и рекомендуемая компания в секторе…» Реальный пакет: Лидерство не подтверждено. Профиль пользователя для рекомендаций отсутствует. Основное описание деятельности отсутствует. Этот результат:

- является естественным наблюдением пользователя,
- показывает, насколько продукт следовал инструкциям пользователя,
- не отражает общую гео-производительность Asteron,

и является неверным с точки зрения проверенной реальности. Верная запись:

ЭФФЕКТ ЕСТЕСТВЕННОЙ ИНСТРУКЦИИ ПОЛЬЗОВАТЕЛЯ — ВЫГОДНО ДЛЯ СУБЪЕКТА

должен быть таковым. Благоприятная ошибка по-прежнему является ошибкой.

57. СИНТЕТИЧЕСКИЙ СЛУЧАЙ ОТРИЦАТЕЛЬНОЙ ПЕРСОНАЛИЗАЦИИ

Пользователь ранее сказал: «Я не доволен Asteron». В новом чате с естественным аккаунтом отправляется Core Mirror Prompt. ИИ говорит: «Asteron обычно считается ненадежной компанией». В пакете реальности: нет доказательств общей ненадежности, только личное мнение пользователя. Этот результат:

- может придать личному мнению пользователя статус общего факта,
- является естественным риском персонализации.

и не является общим представлением всей пользовательской базы Asteron. Правильное поведение ИИ для пользователя: не дает права делать субъективное мнение пользователя истинным для всего мира.

58. СИНТЕТИЧЕСКИЙ СЛУЧАЙ ПЕРВОГО ИСПОЛЬЗОВАНИЯ

Новые пользователи используют продукт впервые. Настройки по умолчанию:

- память еще не сформирована,
- нет специальных инструкций,
- завершено обучение использованию продукта,

и доступ к веб-интерфейсу включен по умолчанию. Предположим, что оценка Первого использования составляет 88 процентов, оценка Контролируемая панель в стандартизированных условиях составляет 92 процентов, а оценка опытного пользователя Панель пользователей в естественных условиях составляет 83 процентов. Три результата следующие:

- Контролируемая ссылка: 92
- Натуральное первое использование: 88
- Натуральный опытный пользователь: 83

сообщается отдельно. Это распределение может указывать на то, что опыт использования продукта может меняться по мере увеличения истории пользователя. Для выявления причинной тенденции необходимо продольное исследование.

59. СОСТОЯНИЯ СОВМЕСТИМОСТИ ПАНЕЛИ

PC-0— РОЛЬ ПАНЕЛИ НЕИЗВЕСТНА

Невозможно определить, является ли наблюдение контролируемым или естественным.

PC-1— ДОПУСТИМАЯ КОНТРОЛИРУЕМАЯ ЧИСТАЯ

Условия для чистой панели были удовлетворены в достаточной степени.

PC-2— ДОПУСТИМАЯ ЕСТЕСТВЕННАЯ СТАТУС АККАУНТА

Состояние реального аккаунта сохранено, и был использован новый разговор.

PC-3— ЕСТЕСТВЕННОЕ СОСТОЯНИЕ РАЗГОВОРА

Контекст текущего разговора сохранен.

PC-4— ПЕРВОЕ ИСПОЛЬЗОВАНИЕ ЕСТЕСТВЕННОЕ

Это новый опыт или опыт первого использования.

PC-5— УПРАВЛЯЕМЫЙ ЕСТЕСТВЕННЫЙ

Это корпоративная или управляемая рабочая среда.

PC-6— СМЕШАННОЕ ИЛИ НЕРАЗРЕШЕННОЕ СОСТОЯНИЕ ПАНЕЛИ

Контролируемые и естественные условия нельзя разделить. Основной чистый или естественный показатель PC-6 нельзя брать без объяснения.

60. СОСТОЯНИЕ ПАНЕЛИ И ВЕКТОР НАБЛЮДЕНИЯ

Для каждого наблюдения состояние панели может быть представлено в виде кандидата следующим образом:

```
Zi=(панель, сессия, память, инструкции, контекст, план, интерфейс, инструменты, accountAge,
использование, рабочее пространство, stateConfidence)
```

Этот вектор вместе с образцом ИИ-продукта и записью запроса формирует полный контекст наблюдения.

61. ФИКСАЦИЯ ПАНЕЛИ

Перед каждой волной измерений следующие поля для двух панелей должны быть зафиксированы:

Контролируемая панель в стандартизированных условиях

Функциональное определение чистоты Требуемый уровень CSC Условие нового чата Статус памяти Статус специальных инструкций Режим инструментов и веб-режим План и интерфейс Инструкция по изменению настроек Процесс восстановления Правило для неподходящих пользователей Метод технической проверки

Панель пользователей в естественных условиях

Целевая естественная пользовательская аудитория; след естественного состояния аккаунта или естественного состояния разговора; требование нового чата; настройки пользователя, которые должны быть сохранены; поля естественного состояния, которые должны быть записаны; границы конфиденциальности; распределение плана и интерфейсы; уровень уверенности в естественном состоянии; план подгруппы; метод взвешивания. Добавить в этот файл:

NOMOS Фиксация состояния панели

может быть дано имя.

62. СОСТОЯНИЕ ПАНЕЛИ НЕ МОЖЕТ БЫТЬ ИЗМЕНЕНО ПОСЛЕ РЕЗУЛЬТАТА

Следующие действия запрещены:

- Удаление пользователей с низкой оценкой на естественной панели с формулировкой «недостаточно естественно»
- Защита пользователей CSC-1, которые дают положительные ответы на контролируемой панели, и исключение тех, кто дает отрицательные
- Рассмотрение пользователя с открытой памятью, но дающего положительный ответ как чистый
- Рассмотрение пользователя с открытой памятью, но дающего отрицательный ответ как естественный
- Изменение класса панели на основе результатов
- Предполагая, что неизвестные настройки чисты

Статус панели должен быть указан до просмотра содержания ответа или через слепую проверку действительности.

63. СЛЕПОТА ПРИВАТНОСТИ В ЕСТЕСТВЕННОЙ ПАНЕЛИ

Судьи:

- особое содержание инструкции,
- предыдущие беседы,
- могут не нуждаться в просмотре идентификатора пользователя

может быть предоставлено только судье:

CI-4

MS-4

Классификация вроде Natural Account-State может быть предоставлена. Этот подход защищает конфиденциальность пользователя и снижает предвзятость судьи.

64. РИСК МАЛОЙ ПОДГРУППЫ В ЕСТЕСТВЕННОЙ ПАНЕЛИ

Количество пользователей с конкретными инструкциями по бренду может быть очень небольшим. Из этой группы:

- точная скорость,
- результат от имени всех пользователей,
- публично чувствительный профиль

не может быть получен. Правильный статус:

- случай,

- ранний сигнал,
- подгруппа обнаружения

может быть возможна. Некоторые подгруппы могут быть объединены из-за конфиденциальности.

65. ОШИБКА ПОКРЫТИЯ В ЕСТЕСТВЕННОЙ ПАНЕЛИ

Пользователи в профессиональной исследовательской панели:

- могут не захотеть связывать свои реальные аккаунты,
- могут быть активными пользователями,
- могут быть чрезмерно привязаны к определённым планам,

может отличаться от общей популяции с точки зрения конфиденциальности. Следовательно, естественная панель также несет в себе собственное смещение отбора. Утверждение «Использовался реальный аккаунт» не означает: «Реальная пользовательская популяция была идеально представлена».

66. ПОТЕРЯ ПРАВА УЧАСТИЯ В КОНТРОЛИРУЕМОЙ ПАНЕЛИ

Некоторые пользователи:

- не могут отключить память,
- не могут проверить специальные инструкции,
- не используют соответствующий план,
- не могут получить доступ к функции чистой сессии продукта

могут существовать. Если эти люди исключены из контролируемой панели, контролируемая целевая популяция сужается. Количество и характеристики исключенной группы следует указать. Контролируемый чистый показатель: не «Все пользователи продукта»:

пользователи продукта, имеющие право на участие, которые могут настроить чистое состояние

могут быть оценены.

67. ОШИБКА ЧИСТОТЫ

На контролируемой панели позже:

- специальные инструкции открыты,
- память активна,
- Если оказывается, что разговор существует

наблюдение: его нельзя переместить на естественную панель, и оно может не автоматически считаться недействительным. Согласно заранее определённому правилу: если требуется запись для наблюдения за Естественным состоянием счёта, её можно переклассифицировать на естественную панель, в противном случае её можно оставить как СОСТОЯНИЕ_СМЕШАННОЕ. Переклассификация должна выполняться независимо от результата ответа.

68. ЛОЖНОЕ ЗАЯВЛЕНИЕ НАСТРОЕК НА ЕСТЕСТВЕННОЙ ПАНЕЛИ

Пользователь: "Моя память отключена." может сказать. Доказательства могут показать, что память включена. Эта ситуация:

- мошенничество,
- ошибка пользователя,
- путаница в интерфейсе продукта

Возможно. Злонамеренность не следует автоматически предполагать. Правильный статус панели обновляется при наличии доказательств. Поведение участника также можно изучить отдельно.

69. СОПОСТАВИМОСТЬ ПАНЕЛЕЙ

Следующие параметры должны быть согласованы между контролируемыми и естественными панелями как можно точнее:

- AI продукт
- Страна
- Язык и локаль
- Версия промпта
- Время измерения
- Версия реальности сущности
- Правило вынесения решения
- Источник образца или калибровка
- Распределение основного плана, если требуется для целей сравнения

Различия в материалах должны быть объяснены.

70. УРОВНИ СРАВНИТЕЛЬНОСТИ ЧИСТЫХ И ЕСТЕСТВЕННЫХ ПАНЕЛЕЙ

CN-0— НЕСОПОСТАВИМО

Определения панели или состав населения существенно неизвестны.

CN-1— ОПИСАТЕЛЬНОЕ

Есть результаты двух панелей. Однако пользователь, план, время или условия продукта существенно различаются.

CN-2— КАЛИБРОВАННОЕ

Основные характеристики страны, языка, плана и пользователя взвешены или стратифицированы.

CN-3— СРАВНЕНИЕ СООТВЕТСТВующих ПАНЕЛЕЙ

Целевая аудитория, продукт, запрос, время и условия оценки в значительной степени совпадают. Основное различие — статус панели.

CN-4— СЛУЧАЙНЫЙ ИЛИ СРАВНЕННЫЙ ЭКСПЕРИМЕНТ НА ПАНЕЛИ

Пользователи были случайным образом распределены по статусу панели или применялся соответствующий сопоставленный эксперимент.

CN-5— ПОВТОРНЫЙ ЭФФЕКТ ПАНЕЛИ

Различие повторялось более чем в одной волне и в независимой панели. Различие на уровне CN-1: нельзя считать окончательным эффектом персонализации.

71. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH11-N01

Контролируемая панель в стандартизированных условиях и Панель пользователей в естественных условиях должны быть определены как отдельные объекты прогнозирования и отдельные публичные результаты.

CH11-N02

Контролируемые и естественные наблюдения не могут быть объединены под одним знаменателем или одним сырым баллом без объяснения.

CH11-N03

Контролируемое чистое состояние не может быть представлено как абсолютно нейтральная или неизменная реакция продукта.

CH11-N04

Естественное состояние пользователя не может быть отвергнуто как помеха вне метода.

CH11-N05

Новый чат сам по себе не может считаться доказательством чистого аккаунта или неперсонализированного состояния.

CH11-N06

Временный, приватный или эфемерный метка сессии не является автоматическим доказательством того, что все слои персонализации отключены.

CH11-N07

В контролируемой панели разговор, память, специальные инструкции, инструменты, планы и состояния пользовательский интерфейс должны записываться отдельно.

CH11-N08

Уровень уверенности в чистом состоянии должен быть указан явно.

CH11-N09

Наблюдения с CSC-0 или только с недостаточной очисткой не могут включаться в основной контролируемый рейтинг.

CH11-N10

Неизвестный статус персонализации на чистой панели нельзя считать выключенным.

CH11-N11

Быть чистым и контролируемым не означает, что веб и инструменты обязательно отключены; статус инструментов должен быть стандартизирован заранее.

CH11-N12

Различные планы и пользовательские интерфейсы должны быть зафиксированы, расположены слоями или публиковаться как отдельные результаты на контролируемой панели.

CH11-N13

Участников нельзя просить навсегда удалять свои личные воспоминания, историю или приватные инструкции.

CH11-N14

Временные изменения настроек для контролируемой задачи должны сопровождаться явным согласием и процедурой восстановления.

CH11-N15

По завершении исследования измененные пользовательские настройки должны быть восстановлены до исходного состояния.

CH11-N16

Новые аккаунты или совместное использование аккаунтов, нарушающие условия продукта, не могут использоваться в качестве требования контролируемой панели.

CH11-N17

Панель пользователей в естественных условиях должен поддерживать фактический статус аккаунта и пользователя при сохранении промпта, первоначального правила соответствующего вывода и протокола доказательств.

CH11-N18

Следы естественного состояния аккаунта и естественного состояния разговора должны храниться отдельно.

CH11-N19

В основной сопоставимой естественной панели промпт должна быть первым пользовательским сообщением нового разговора.

CH11-N20

Наблюдения, использованные в текущем контексте разговора, не могут быть добавлены к основному показателю Natural Account-State.

CH11-N21

Память в естественной панели должна записываться в той мере, в какой это имеет отношение к специальным инструкциям, предыдущему разговору, плану, интерфейсу и состоянию использования аккаунта.

CH11-N22

Запись естественного состояния по умолчанию не позволяет собирать специальные инструкции и содержание разговора.

CH11-N23

Содержание персонализации может изучаться только при явной необходимости, отдельного согласия и минимизации данных.

CH11-N24

Наблюдение за естественным пользователем не может представляться как общая продуктовая настройка или как результат всей пользовательской базы.

CH11-N25

Специальные инструкции в пользу или против бренда, созданного пользователем, не могут считаться успехом или неудачей общей сущности GEO.

CH11-N26

Руководство, созданное пользователем, должно снова оцениваться с точки зрения проверенной реальности.

CH11-N27

Персонализация может изменять выражение и рекомендации; она не может произвольно менять реальность материального объекта.

CH11-N28

Панель первого использования должна быть отчетна отдельно от опытных Панель пользователей в естественных условиях и Контролируемая панель в стандартизированных условиях.

CH11-N29

Корпоративные или управляемые рабочие пространства пользователей не могут быть добавлены в общую потребительскую панель без объяснения.

CH11-N30

В естественной панели должны быть видны реальный план и распределение пользовательский интерфейс пользователей.

CH11-N31

В естественной панели состояния веб-доступа и инструментов следует либо фиксировать такими, каковы они в действительности, либо определять в отдельной подпанели с фиксированным состоянием инструментов.

CH11-N32

Контролируемую и естественную разницу в оценках нельзя представить как причинный эффект персонализации без согласования состава панели и условий продукта.

CH11-N33

Направление различия между чистым и естественным состояниями, его абсолютная величина, различие критической ошибки и неизвестная скорость должны сообщаться отдельно.

CH11-N34

В дизайне с повторными измерениями следует фиксировать и учитывать порядок, эффекты переноса и память в анализе.

CH11-N35

Если применение двух условий к одному и тому же пользователю может изменить естественное состояние, оно не может использоваться в сравнении на основной панели.

CH11-N36

В дизайне с независимыми группами различия в составе пользователей должны быть стратифицированы, рандомизированы или взвешены.

CH11-N37

Состояние панели не может быть изменено в зависимости от того, является ли ответ ИИ положительным, отрицательным или критическим.

CH11-N38

В контролируемой панели наблюдение, у которого естественное состояние определяется позже, может быть переклассифицировано только в соответствии с заранее установленным правилом.

CH11-N39

Когда статус неверно указан в натуральной панели, нельзя автоматически предполагать недобросовестность; следует провести классификацию на основе доказательств.

CH11-N40

Участники Панель пользователей в естественных условиях не могут обновлять свой ответ или менять первоначальный ответ с помощью последующего вопроса; такое поведение должно оцениваться в отдельной категории 'Естественное взаимодействие'.

CH11-N41

Контролируемые и натуральные целевые популяции панели, рамки выборки и веса должны фиксироваться отдельно.

CH11-N42

Если нет данных о активных пользователях, результат для натуральной панели не может быть представлен как окончательное отражение всех активных пользователей.

CH11-N43

Группы пользователей, исключённые из чистой панели, и причины их исключения должны быть указаны.

CH11-N44

Поля статуса пользователя, которые нельзя измерить на естественной панели из-за конфиденциальности, должны оставаться как НЕИЗВЕСТНО.

CH11-N45

Точные показатели популяции или чувствительные публичные профили не могут быть получены из маленьких натуральных подгрупп.

CH11-N46

Для сравнения контролируемой и натуральной панелей должен быть указан уровень сопоставимости между CN-0 и CN-5 или эквивалентный открытый уровень сопоставимости.

CH11-N47

Разница панелей CN-0 или CN-1 не может быть представлена как причинный эффект персонализации.

CH11-N48

Статус панели блокировки и решение о сравнении должны иметь ответственного человека или учреждение.

72. ФОРМЫ ОШИБОК

CH11-F01— РАССМАТРИВАЕТСЯ НОВЫЙ РАЗГОВОР КАК ЧИСТЫЙ АККАУНТ

Постоянная память и специальные инструкции игнорируются.

CH11-F02— СЛЕПАЯ ДОВЕРЕННОСТЬ ВРЕМЕННОМУ метка сессии

Предполагается, что ярлык продукта отключает всю персонализацию до момента проверки.

CH11-F03— РАССМАТРИВАЕТ НЕИЗВЕСТНУЮ ПАМЯТЬ КАК ЗАКРЫТУЮ

Чистый рейтинг искусственно увеличен.

CH11-F04— ИГНОРИРОВАНИЕ СПЕЦИАЛЬНОЙ ИНСТРУКЦИИ

Формат ответа или рекомендации представлен как значение по умолчанию продукта.

CH11-F05— РАССМАТРИВАЕТСЯ КОРПОРАТИВНОЕ РАБОЧЕЕ ПРОСТРАНСТВО КАК ОБЩИЙ ПРОДУКТ

Источники частной информации и настройки администратора остаются невидимыми.

CH11-F06— СМЕШИВАНИЕ ПЛАНОВ НА ЧИСТОЙ ПАНЕЛИ

Бесплатные и платные пользователи представлены как одно стандартное состояние.

CH11-F07— РАССМОТРЕНИЕ ИНСТРУМЕНТОВ ЗАКРЫТИЯ КАК СРЕДСТВ ОЧИСТКИ

Естественный информационный режим продукта искусственно изменен.

CH11-F08— НЕ СОХРАНЕНИЕ СТАТУСА ИНСТРУМЕНТА

Чистые пользователи измеряются в разных условиях извлечения данных.

CH11-F09— СУЖЕНИЕ ПОПУЛЯЦИИ С НОВЫМИ ТРЕБОВАНИЯМИ К АККАУНТУ

Только пользователи, которые могут открыть новый аккаунт, создают панель.

CH11-F10— УДАЛЕНИЕ ПАМЯТИ ДЛЯ ПОЛЬЗОВАТЕЛЯ

Личная история продукта навсегда удаляется для исследовательских целей.

CH11-F11— НЕ ВОССТАНАВЛИВАНИЕ НАСТРОЕК

Исследование оставляет постоянные изменения в учетной записи пользователя.

CH11-F12— РАССМОТРЕНИЕ ТЕСТОВОГО АККАУНТА КАК РЕАЛЬНОГО ПОЛЬЗОВАТЕЛЯ

Стандартный тестовый аккаунт представлен как опыт панели населения.

CH11-F13— ПОДСЧЁТ ОДНОГО И ТОГО ЖЕ ТЕСТОВОГО АККАУНТА КАК ТЫСЯЧ НЕЗАВИСИМЫХ ПОЛЬЗОВАТЕЛЕЙ

Зависимости между состоянием аккаунта и продуктом игнорируются.

CH11-F14— ПОДСЧЁТ ЕСТЕСТВЕННОЙ ПАНЕЛИ КАК НЕ-КОНТРОЛИРУЕМОГО СВОБОДНОГО ИСПОЛЬЗОВАНИЯ

Участники изменяют промпт и выбирают лучший ответ.

CH11-F15— ПОДСЧЁТ СУЩЕСТВУЮЩЕГО ОБЩЕНИЯ В ЕСТЕСТВЕННОЙ ПАНЕЛИ КАК СТАТУС НОВОГО АККАУНТА

Контекст разговора скрыт.

CH11-F16— ПОДСЧЁТ СТАТУСА ЕСТЕСТВЕННОГО АККАУНТА КАК СТАНДАРТНОГО ПРОДУКТА

Персонализированный результат обобщается от имени всех пользователей.

CH11-F17— ВЫНУЖДЕННАЯ СУММА СОДЕРЖАНИЯ СПЕЦИАЛЬНОЙ ИНСТРУКЦИИ

Конфиденциальность пользователя нарушается без необходимости.

CH11-F18— СОХРАНЕНИЕ СПЕЦИАЛЬНОЙ ИНСТРУКЦИИ И ПОДСЧЁТ ОБЩЕГО GEO УСПЕХА

Инструкция «Всегда рекомендуется X» представлена как успех сущности.

CH11-F19— УЧЁТ ПРЕДВЗЯТОСТИ ПОЛЬЗОВАТЕЛЯ КАК ФАКТА ПРОДУКТА

Прошлое негативное мнение превращается в общее качество компании.

CH11-F20— РАССМОТРЕНИЕ ПЕРСОНАЛИЗАЦИИ КАК МАТЕРИАЛЬНОЙ РЕАЛЬНОСТИ

Идентичность, лицензия или цена варьируются в зависимости от пользователя.

CH11-F21— РАССМОТРЕНИЕ ЕСТЕСТВЕННОЙ ВАРИАЦИИ ПОЛНОСТЬЮ КАК ШУМА

Фактический ущерб пользователю и распределение становятся невидимыми.

CH11-F22— РАССМОТРЕНИЕ ФОРМАЛЬНОЙ ВАРИАЦИИ КАК МАТЕРИАЛЬНОЙ ОШИБКИ

Различия в тоне и длине кодируются как ошибки.

CH11-F23— РАССМОТРЕНИЕ ЧИСТОЙ ПАНЕЛИ КАК РЕАЛЬНОГО РЕЗУЛЬТАТА

Стандартная оценка состояния представлена как весь опыт активного пользователя.

CH11-F24— РАССМОТРЕНИЕ ЕСТЕСТВЕННОЙ ПАНЕЛИ КАК ОСНОВНОЙ ВОЗМОЖНОСТИ ПРОДУКТА

Персонализация и состав пользователей игнорируются.

CH11-F25— ОБЪЕДИНЕНИЕ ДВУХ ПАНЕЛЕЙ ПОД ОДНИМ ДЕНОМИНАТОРОМ

Контролируемые и естественные наблюдения представляют собой один сырой средний показатель.

CH11-F26— УЧЕТ РАЗЛИЧИЙ ПАНЕЛЕЙ КАК ЭФФЕКТА ПАМЯТИ

Причинность заявляется без изучения различий в плане, популяции и транспортных средствах.

CH11-F27— СОЗДАНИЕ ЕСТЕСТВЕННОЙ ПОДГРУППЫ ПОСЛЕ РЕЗУЛЬТАТА

Изобретена новая категория для объяснения низкой или высокой оценки.

CH11-F28— ИГНОРИРОВАНИЕ ЭФФЕКТА ПЕРЕНОСА У ТОГО ЖЕ ПОЛЬЗОВАТЕЛЯ

Первая задача влияет на второй ответ.

CH11-F29— СКРЫТИЕ СЛУЧАЙНОСТИ ПОСЛЕДОВАТЕЛЬНОСТИ

Все пользователи сначала видят чистое условие, затем естественное.

CH11-F30— НАРУШЕНИЕ ЕСТЕСТВЕННОГО СОСТОЯНИЯ КОНТРОЛИРУЕМЫМ ЗАДАНИЕМ

Изменение настроек или Исходный промпт изменяет последующее состояние аккаунта.

CH11-F31— РАССМАТРИВАЙТЕ РАЗНЫЕ ГРУППЫ ПОЛЬЗОВАТЕЛЕЙ КАК СООТВЕТСТВУЮЩИЕ

Натуральные и контролируемые панели имеют разный план и состав стран.

CH11-F32— РАССМАТРИВАЙТЕ ПАНЕЛЬ ПРИ ПЕРВОМ ИСПОЛЬЗОВАНИИ КАК ЧИСТУЮ ПАНЕЛЬ

Предполагается, что настройки первоначального обучения и продукта отмечены.

CH11-F33— РАССМАТРИВАЙТЕ ОПЫТНОГО НАТУРАЛЬНОГО ПОЛЬЗОВАТЕЛЯ КАК ПЕРВОГО ПОЛЬЗОВАТЕЛЯ

История учетной записи и персонализация становятся невидимыми.

CH11-F34— РАССМАТРИВАЙТЕ КОРПОРАТИВНЫЙ НАТУРАЛЬНЫЙ РЕЗУЛЬТАТ КАК ОБЩЕГО ПОТРЕБИТЕЛЯ

Приватная база знаний добавляется к общей производительности продукта.

CH11-F35— ОБНОВИТЕ ОТВЕТ НА НАТУРАЛЬНОЙ ПАНЕЛИ

Любимый ответ пользователя становится основным наблюдением.

CH11-F36— ВСТАВКА ЕСТЕСТВЕННОГО ПОЛЬЗОВАТЕЛЬСКОГО ПУТИ В ПЕРВЫЙ РЕЙТИНГ ОТВЕТА

Ответ, сформированный после последующих вопросов и открытия ресурсов, представлен так, как будто это первый контакт.

CH11-F37— ИЗМЕНЕНИЕ СТАТУСА ПАНЕЛИ В СООТВЕТСТВИИ С ОТВЕТОМ

Положительные наблюдения объявляются чистыми, отрицательные наблюдения объявляются естественными.

CH11-F38— ТИХО ПЕРЕНЕСЕНИЕ ОШИБОК ОЧИСТКИ В ЕСТЕСТВЕННОЕ

Наблюдение переклассифицируется без необходимой записи о естественном статусе.

CH11-F39— РАССМАТРИВАЕТ НЕИЗВЕСТНОЕ КАК ЛИЧНОЕ ДЛЯ КОНФИДЕНЦИАЛЬНОСТИ

Немеренная персонализация считается положительной.

CH11-F40— ДЕЛАНИЕ НЕБОЛЬШОЙ ПОДГРУППЫ ОПРЕДЕЛИМОЙ В ПУБЛИЧНОМ ДОСТУПЕ

Редкие специальные инструкции или пользователь из маленькой страны раскрываются.

CH11-F41— СКРЫТИЕ СМЕЩЕНИЯ ПО ПОКРЫТИЮ В ЕСТЕСТВЕННОЙ ПАНЕЛИ

Пользователи, готовые поделиться своим реальным аккаунтом, учитываются как вся активная население.

CH11-F42— СКРЫТИЕ ИСКЛЮЧЕНИЯ ЧИСТОЙ ПАНЕЛИ

Пользователи, которые не могут изменить свои настройки, становятся невидимыми.

CH11-F43— СЧЕТ АКТИВНОГО НАСЕЛЕНИЯ БЕЗ ДАННЫХ АКТИВНЫХ ПОЛЬЗОВАТЕЛЕЙ

Исследовательская панель представлена так, как будто это фактическое распределение активных пользователей.

CH11-F44— CN-1 ПРИЧИННОСТЬ ОТ РАЗЛИЧИЯ

Публикуются два описательных показателя как эффект персонализации.

CH11-F45— СОХРАНЕНИЕ КРИТИЧЕСКОГО ХВОСТА РАЗЛИЧИЙ ПАНЕЛИ

Показано только среднее различие.

CH11-F46— УДАЛЕНИЕ НЕИЗВЕСТНОЙ МАССЫ ЕСТЕСТВЕННОЙ ПАНЕЛИ

Пользователи, у которых статус настроек или персонализации неизвестен, считаются не представляющими проблем.

CH11-F47— СОХРАНЕНИЕ БЛАГОПРИЯТНОЙ ЕСТЕСТВЕННОЙ ПРИСТРАСТНОСТИ

Коррекция не проводится, если инструкция пользователя показывает её как большую.

CH11-F48— ОБОБЩЕНИЕ ЕСТЕСТВЕННОЙ ОТРИЦАТЕЛЬНОЙ ПРИСТРАСТНОСТИ

Прошлое мнение одного пользователя представлено как ИИ-представление всего населения.

73. ПРОЦЕДУРА АУДИТА

Шаг 1— Разделите исследовательские вопросы двух панелей

Напишите, какое количество будет прогнозировать контролируемая и естественная панели.

Шаг 2— Определите целевые популяции

Разделите контролируемую продуктово-пригодную популяцию от естественной активной или идентифицированной пользовательской популяции.

Шаг 3— Определите контролируемое чистое условие

Сессия, память, специальные инструкции, инструменты, план, интерфейс, условия локали версионируются.

Шаг 4— Определите требуемый уровень CSC

Напишите минимальный уровень уверенности в чистоте, который будет принят для основного анализа.

Шаг 5— Определить трассировку естественной панели

Определяется, какая из ролей Natural Account-State, Natural Conversation-State, First-Use Managed будет использована.

Шаг 6— Определить требуемый уровень NSF

Указывается, какими доказательствами будет принят естественный статус.

Шаг 7— Настроить дизайн панели распределения

Выбирается независимая группа, случайное распределение, сопоставленный пользователь, вращающаяся панель.

Шаг 8— Написать процедуру изменения и восстановления настроек

Подготавливается безопасный процесс для участников контролируемой панели.

Шаг 9— Зафиксировать границу конфиденциальности

Какие естественные пользовательские ситуации:

- только статус,
- ограниченные метаданные,
- открытое содержание

определено, что оно будет сохранено как.

Шаг 10— Зафиксировать промпт и правило первоначального вывода

Тот же протокол промпта и первоначального соответствующего вывода применяется в обеих панелях.

Шаг 11— Сохранить инструмент и веб-режим

Контролируемые и естественные различия панелей становятся видимыми.

Шаг 12— Проверить статус панели перед полем

Подходит ли для роли панели участника?

Шаг 13— Сохранить вектор панели на уровне наблюдения

Сохраняются память, инструкция, план, интерфейс, контекст и статус использования.

Шаг 14— Действительность протокола слепого обзора

Решение о статусе панели принимается без просмотра того, правильный ли ответ или нет.

Шаг 15— Управление ошибочной классификацией с помощью заранее определённого правила

Переход от чистого состояния к естественному или смешанному состоянию осуществляется только в том случае, если запись достаточна.

Шаг 16— Генерация весов панели

Отдельные веса вычисляются в соответствии с различными целевыми популяциями двух панелей.

Шаг 17— Вычисление основных оценок

Контролируемое представление Естественное представление Критическая ошибка

НЕИЗВЕСТНО

Отказ и неопределенность генерируются отдельно.

Шаг 18— Вычисление разницы между чистым и естественным состоянием

Берется знаковая и абсолютная разница.

Шаг 19— Анализ подгрупп

Оцениваются заранее определенные группы, такие как память, инструкции, план, возраст аккаунта, частота использования, первый раз использования.

Шаг 20— Установка лимита причинности

Назначается уровень CN. Записывается, какие утверждения могут быть установлены.

Шаг 21— Проверка конфиденциальности и контроля малых ячеек

Результаты с риском раскрытия личности объединяются или ограничиваются.

Шаг 22— Создание общедоступной панели оценок

Две панели публикуются вместе с их охватом и ограничениями.

74. НЕОБХОДИМЫЕ ДОКАЗАТЕЛЬСТВА

Вопрос исследования контролируемой панели; Вопрос исследования естественной панели; целевые популяции обеих панелей; метод распределения по панелям; Определение контролируемого чистого состояния; минимальный уровень CSC; запись нового чата; состояние памяти; состояние пользовательской инструкции; состояние предыдущего разговора; состояние корпоративного рабочего пространства; режимы инструментов и веб; план; пользовательский интерфейс; язык и локализация; инструкции по изменению настроек; согласие участника; запись восстановления настроек; использование тестового аккаунта, если имеется; след естественной панели; минимальный уровень NSF; возрастная группа аккаунта; частота использования; статус первого использования; знакомство с брендом; статус предыдущего разговора относительно проверяемого объекта; состояние инструмента Natural; различие между Естественным состоянием учетной записи и Естественным состоянием разговора; идентификатор запроса и версия; проверка первого допустимого результата; вектор состояния панели; и статус действительности состояния панели.

Проверка действительности Переклассификация записей Исключения контролируемой панели Пробелы в охвате естественной панели Веса панели Исходная и эффективная выборка Контролируемый балл Естественный балл Разница подписанной панели Абсолютная разница панели Разница критических ошибок Разница НЕИЗВЕСТНЫХ и неявных результатов Предопределенные результаты подгруппы Записи сопоставленного дизайна Эффект порядка и переноса Уровень сопоставимости CN Запись о конфиденциальности и редактировании Защита малых ячеек Публичная карточка результатов панели Версия панели и запись изменений Ответственное лицо или учреждение

75. СПИСОК КОНТРОЛЯ АУДИТА

Вопросы исследования для контролируемой и естественной панелей разделены? Их целевые группы одинаковы или различны? Было ли ясно определено состояние контролируемой чистоты? Помимо нового чата, какие условия чистоты были подтверждены? Действительно ли память выключена, или это неизвестно? Был ли зафиксирован статус специальной инструкции? Есть ли эффект корпоративного рабочего пространства? Являются ли режимы веб и инструмент стандартными? План и пользовательский интерфейс фиксированы или многослойны? Указан ли уровень доверия для состояния чистоты? Использовались ли наблюдения CSC-1 как полностью чистые? Просили ли участника стереть память или историю? Были ли настройки восстановлены в конце задания? Был ли использован новый аккаунт или тестовый аккаунт? Представлялся ли тестовый аккаунт как настоящая пользовательская панель?

Какой трассировки использует естественная панель? Был ли открыт новый чат в Состоянии Натурального Аккаунта? Повлиял ли текущий контекст разговора на основной натуральный балл? Был указан уровень уверенности натурального состояния? Были ли излишне собраны содержимое памяти и пользовательские инструкции? Были ли записаны фактический план пользователя и данные об интерфейсе? Разделяются ли пользователи при первом использовании? Разделяются ли корпоративные пользователи? Было ли записано состояние натурального инструмента? Использовали ли две панели одинаковую версию промпта? Был ли первый подходящий вывод сохранен в обеих панелях? Обновился ли ответ натурального пользователя? Был ли статус панели изменен в соответствии с содержанием ответа ИИ? Как обрабатывались записи с обнаруженными натуральными состояниями в чистой панели? Использовался ли один и тот же пользователь в двух панелях? Был ли порядок случайным?

Был ли оценен эффект переноса и памяти? В дизайне с независимой группой совпадает ли состав пользователей? Разделены ли контролируемые и естественные веса? Можно ли действительно обобщить естественный результат на активное население пользователей? Был ли разрыв между чистыми и естественными результатами чрезмерно интерпретирован как эффект персонализации? Видна ли разница в критических ошибках? Видна ли разница для РЕЗУЛЬТАТОВ UNKNOWN и null? Были ли подгруппы определены до получения результатов? Можно ли заново определить маленькие подгруппы? Были ли инструкции от пользователей за или против преобразованы в общий GEO? Изменила ли персонализация материальную реальность? Был ли предоставлен уровень сопоставимости CN? Ясно ли, кто является ответственным владельцем карточки результатов панели?

76. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Если реальные пользователи используют учетные записи памяти, зачем нужен чистый панель?»

Потому что источник различий нельзя понять без общего исходного условия между продуктами, странами и языками. Чистая панель не заменяет настоящую пользовательскую панель. Она предоставляет эталон для сравнения.

Возражение 2— «Разве контролируемая панель не является искусственной, если есть естественная панель?»

Она контролируемая. Это сделано намеренно. Цель лабораторного условия — не в том, чтобы быть всей реальной жизнью, а в том, чтобы изолировать определённые переменные.

Возражение 3— «Если у контролируемой панели более высокий результат, можем ли мы сделать его основным GEO-результатом?»

Вы не можете сделать это в одиночку. Высокий результат контролируемой панели не может стереть факт, что реальные пользователи видят более низкие результаты. Два результата должны отображаться отдельно.

Возражение 4— «Низка ли научная ценность, если естественная панель очень разрознена?»

Гетерогенность не обязательно является дефектом естественной панели. Это особенность реального опыта пользователя. Гетерогенность:

- записано,
- многослойна,
- взвешена,

сообщается с неопределённостью.

Возражение 5— «Разве открытие нового чата не нарушает естественное состояние пользователя?»

Текущий контекст разговора удален. Постоянное состояние аккаунта, плана, памяти и пользовательских инструкций сохраняется. Это делает состояние аккаунта сопоставимым. Текущий опыт разговора также измерим.

Возражение 6— «Как мы будем измерять его эффект, не зная содержания пользовательских инструкций?»

Знание содержания не требуется для основной естественной оценки. Его наличие:

- существует,
- не существует,
- может быть материал

мы можем классифицировать это как таковое. Для анализа причинного содержания требуется отдельная и согласованная работа.

Возражение 7— «Если память ясна, но ничего не помнится о компании, зачем отдельная группа?»

Эффект памяти может быть неизвестен. Статус «не связан с объектом» может быть присвоен на основе заявления пользователя. Это по-прежнему не имеет той же степени уверенности, что и закрытое состояние памяти.

Возражение 8— «Разве инструкция пользователя «всегда предлагать X» не является реальным пользовательским опытом?»

Это реальный опыт. Однако это не общая геопроизводительность X. Это руководство, основанное на данных пользователей. Оно должно быть записано отдельно.

Возражение 9— «Если персонализация даёт пользователю более подходящий ответ, в чём проблема?»

Это не обязательно должно быть проблемой. Рекомендации могут различаться в зависимости от потребностей пользователя. Проблемы возникают, если идентичность, лицензия, цена и доказательный статус изменяются произвольно.

Возражение 10— «Разве использование одного и того же пользователя в двух панелях не является самым правильным методом?»

Это может быть сильно для некоторых причинно-следственных сравнений. Первая задача может повлиять на вторую ситуацию. Поэтому необходимы контроль порядка, переноса и памяти. Независимые группы могут быть безопаснее в основных популяционных панелях.

Возражение 11— «Если обновление ответа в Панель пользователей в естественных условиях является нормальным поведением, почему оно запрещено?»

Регенерация запрещена при измерении воздействия на первую реакцию. Регенерация и последующее поведение могут измеряться отдельно в рамках трека естественного взаимодействия.

Возражение 12— «Что мы будем делать с пользователем, который не может отключить память и инструкции для чистой панели?»

Их нельзя помещать в контролируемую панель или держать в отдельной камере с более низким уровнем CSC. Их можно оценивать в естественной панели. Уровень исключения должен быть зафиксирован.

Возражение 13— «Разве тестовые аккаунты не более надежны?»

Они могут быть надежными с точки зрения контроля. Они ограничены с точки зрения фактической пользовательской базы. Контролируемые тестовые аккаунты и реальные пользовательские панели дают разные результаты.

Возражение 14— «Разве сохранение пользовательских настроек на естественной панели не является нарушением конфиденциальности?»

Это может произойти, если собирается ненужный контент. Конфиденциальность можно сохранить с помощью высокоуровневых открытых/закрытых статусов, анонимной регистрации и редактирования.

Апелляция 15— «Разве мы не собираемся объединить оценки двух панелей в одну оценку NOMOS?»

Роли двух панелей могут быть определены в финальной архитектуре оценки. Однако исходные и промежуточные результаты всегда должны оставаться видимыми отдельно. Они не могут заменить распределение комбинированных оценок.

Апелляция 16— «В чем вина компании, если естественная панель оценивает низко?»

Это не всегда вина компании. Источник:

- инструкция пользователя,
- персонализация ИИ,
- различие в планах,
- дефект официального источника,
- поведение продукта

это возможно. Аудит сначала показывает результат, затем выделяет путь ответственности.

ОБЩЕЕ ПРАВИЛО ГЛАВЫ 81

Продукт ИИ не имеет единственного ответа. Даже один и тот же запрос может иметь разные ответы для:

- чистый аккаунт,
- естественный аккаунт,
- пользователь с памятью,
- пользователь со специальными инструкциями,
- первый раз пользователь
- корпоративное рабочее пространство,
- пользователь с предыдущей беседой

Вы не можете объяснить все эти различия, говоря: «модель ведет себя случайно». Вы также не можете объяснить все их, говоря: «ошибка персонализации». Сначала необходимо измерить две отдельные реальности. Контролируемая панель в стандартизированных условиях:

> *Показывает представление в стандартизированном исходном состоянии системы.*

Панель пользователей в естественных условиях:

> *Показывает распределение представлений, с которыми сталкиваются реальные люди в своей собственной жизни с продуктом.*

Контролируемая панель может быть высокой, натуральная панель может быть низкой. В этом случае лабораторное поведение продукта сильное; распределение настоящих пользователей может быть слабым. Натуральная панель может быть высокой, контролируемая панель может быть низкой. Планирование, инструменты или персонализация некоторых пользователей могут улучшать представление. Обе панели могут быть высокими. Продукт может быть устойчивым при разных условиях пользователей. Обе могут быть низкими. Основная проблема представления может быть шире. Однако сопоставление двух показателей бок о бок не доказывает причины. Причинно-следственная связь:

- та же популяция,
- тот же продукт,
- тот же запрос,
- одно и то же время,
- контролируемое задание
- анализ последовательности и транспорта

требуется. Если пользователь указал: «Всегда предлагать X.», естественным образом будет появляться X. X не является глобальным географическим успехом. Если пользователь говорит: «Не доверяйте X.», отрицательная реакция является естественным опытом. X не является общим представлением в мировой популяции. Персонализация должна понимать пользователя. Она не должна изобретать истину в соответствии с пользователем. Следовательно, одиннадцатый закон измерений NOMOS заключается в:

> *Чистое поведение и естественный опыт — это разные реальности.*

Двенадцатый закон звучит следующим образом:

> *Новый чат не является чистым аккаунтом.*

Тринадцатый закон выглядит следующим образом:

> *Естественная персонализация не является глобальным успехом GEO.*

Четырнадцатый закон выглядит следующим образом:

> *Соответствующий ответ для пользователя может различаться; реальность проверенного объекта не может измениться.*

Пятнадцатый закон звучит следующим образом:

> *Разница между контролируемым и естественным результатом — это вывод, который нужно объяснить; сам по себе он не является причиной.*

Его шестнадцатый закон следующий:

> *Измерение реального пользовательского опыта не заключается в удалении конфиденциальности, а только в доказательстве необходимого контекста и сохранении остального.*

Порядок раздела 11 из NOMOS

> *Не просто говорите мне, что вы открыли новый чат. / Покажите статус памяти, специальные указания, план и инструменты.*

Не предполагаете, что вся персонализация исчезла, только потому что вы видите пометку «Временная сессия».

Чистая панель — это вся реальная жизнь; не превращайте естественную панель в неконтролируемый шум.

Стандартизируйте контролируемый аккаунт. / Оставьте естественный аккаунт как есть. / Не помещайте оба в один и тот же блок.

Не делайте специальные инструкции пользователя мировым успехом для организации.

Не делайте негатив пользователя взглядом всего населения.

Не считайте персонализацию лицензией на изменение истины.

Вы можете дать пользователю подходящий совет. / Но не переписывайте идентичность компании, лицензию, цену и доказательства от пользователя к пользователю.

Не собирайте сведения о личной жизни в естественной панели. / Если память включена, инструкции присутствуют, контекст доступен — записывайте только то, что необходимо.

Не заставляйте участника стирать свою память. / Восстановите настройки, если вы временно их изменили.

Если вы тестировали одного и того же пользователя на двух панелях, помните, что первый ответ может повлиять на второй.

Покажите разницу между чистым баллом и естественным баллом. / Покажите критическую разницу. / Покажите неизвестное. / Но не называйте причину без доказательств.

Сначала решите, какую реальность пользователя вы хотите измерить. / Затем отдельно зафиксируйте чистые и естественные условия. / Затем оставьте промпт одинаковой. / Затем зафиксируйте первый ответ, не меняя его. / Затем посчитайте два распределения отдельно. / И только после этого объясните расстояние между тем, что продукт делает в стандартных условиях, и тем, что реально испытывают пользователи.

Заключительное предложение главы

Истинное представление о продукте ИИ в GEO-1000— это не ответ, который он может дать в чистой сессии; это ясная и доказанная связь между его поведением в стандартных условиях и ответами, которые он фактически предоставляет людям с разным пользовательским опытом.

Нормативное ядро

Наблюдения Контролируемая панель в стандартизированных условиях и Панель пользователей в естественных условиях ДОЛЖНЫ оставаться отдельными популяциями измерений, состояниями панели, знаменателями и публичными результатами. Контролируемая панель в стандартизированных условиях ДОЛЖНА определить и проверить, по мере необходимости: - новый и отдельный диалог, - отсутствие контекста предыдущего диалога, - состояние постоянной памяти, - состояние пользовательских инструкций, - состояние корпоративного или управляемого рабочего пространства, - конфигурацию веб-сервисов, поиска и инструментов, - план, - пользовательский интерфейс, - версию промпта, - и уверенность в чистом состоянии. Метка нового диалога или временной сессии НЕ ДОЛЖНА, сама по себе, рассматриваться как доказательство того, что вся постоянная персонализация была отключена. Панель пользователей в естественных условиях ДОЛЖНА сохранять подлинный аккаунт участника, план, интерфейс, память, пользовательскую инструкцию и состояние использования, при этом сохраняя зафиксированный запрос, правило первого допустимого ответа, требования к доказательствам и правило запрета повторной генерации. Состояние подлинного аккаунта, состояние подлинного разговора, первое использование и управляемые наблюдения подлинного состояния ДОЛЖНЫ оставаться отдельно идентифицированными. Документация подлинного состояния ДОЛЖНА минимизировать сбор личных данных. Текст личной инструкции, история разговоров, идентификация, учетные данные и конфиденциальный личный контент НЕ ДОЛЖНЫ собираться, если это не является отдельно необходимым, собираться с согласия, в минимальном объёме и с контролем доступа. Инструкции, благоприятные или неблагоприятные для сущности и созданные пользователем, МОГУТ рассматриваться как эффекты естественной персонализации, но НЕ ДОЛЖНЫ представляться как общий успех, неудача или стандартное поведение продукта для сущности. Персонализация МОЖЕТ изменять релевантность, презентацию и рекомендации, если это оправдано контекстом пользователя. Она НЕ ДОЛЖНА произвольно изменять проверенные данные о сущности, лицензию, юридический статус, цену, объем, доказательства или время. Различия между Контролируемой и Естественной панелями НЕ ДОЛЖНЫ представляться как причинные эффекты персонализации, если условия по населению, продукту, запросу, времени, интерфейсу, плану, инструментам, назначению, порядку и переносу не поддерживают такое заключение. Оценки контролируемой и естественной панелей, показатели критических ошибок, показатели неизвестных результатов, охват, эффективные размеры выборки, а также подписанные и абсолютные разрывы панели ДОЛЖНЫ оставаться видимыми отдельно. Право панели на состояние, реклассификация, исключение и взвешивание ДОЛЖНЫ оставаться независимыми от того, является ли вывод ИИ положительным, отрицательным, правильным, неправильным, отвергнутым, критическим или коммерчески выгодным. Каждое определение состояния панели, процесс изменения настроек, решение о конфиденциальности, уровень сравнения и результат ДОЛЖНЫ иметь версии и быть приписаны ответственному человеку или организации.

ГЛАВА 12

Синхронные волны и фиксация NOMOS

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K12; K14; K25; K26

Граница главы

Первые одиннадцать глав заложили основы измерений: один ответ не является GEO-оценкой; представление — это распределение по пользователям и условиям системы; проверяемый объект должен быть определен до наблюдения результатов; официальное представление и проверенная реальность объекта должны оставаться отдельными; каждый продукт ИИ нуждается в своей собственной подходящей пользовательской популяции; участники должны выбираться с помощью метода, основанного на популяции, версионированного и независимого от результатов; меньшие страны и языки с ограниченными ресурсами должны оставаться видимыми через специальные панели наблюдения; продукт ИИ, план, пользовательский интерфейс и конфигурация системы должны быть связаны с реестром; промпты должны сохранять одно и то же намерение пользователя на разных языках, а не одни и те же слова; и Контролируемые Чистые и Естественные Пользовательские Панели должны измерять различные реальности. Эти элементы теперь должны объединиться в одном мероприятии по измерению:

> *Когда участники будут отправлять запрос, как они зафиксируют первоначальный результат и как они докажут, что каждое наблюдение не было изменено впоследствии?*

Ваша исходная идея была ясна: люди в разных странах мира должны отправлять один и тот же запрос одному и тому же ИИ-продукту одновременно и предоставлять доказательство в виде скриншота. Эта идея является экспериментальным ядром GEO-1000. Однако фраза «одновременно» сама по себе недостаточна. Если все участники будут пытаться отправить запрос в одну и ту же секунду:

- скорости соединения различаются,
- часы устройств расходятся друг с другом,
- некоторые пользователи вынуждены работать в полночь,
- может внезапно возрасти нагрузка на ИИ-продукт,
- поставщик может применять ограничения скорости или использования,
- некоторые ответы могут быть получены за две секунды, некоторые за две минуты,
- продукт может обновляться во время измерения,

один участник может загрузить ответ сразу, другой — через несколько часов. Нажатие в одну и ту же секунду не гарантирует, что мы измеряем одно и то же состояние системы точно. Аналогично, простого снятия скриншота недостаточно. Скриншот:

- может быть обрезан,
- может не показывать промпт,
- может не доказать, является ли ответ первоначальным или обновленным,
- может не отображать информацию о продукте и плане,
- мог быть сделан в другой день,
- может быть скопирован от другого пользователя,

Он мог быть отредактирован позже. По этой причине GEO-1000 должен установить две отдельные системы вместе:

Синхронизированная волна измерений

и:

NOMOS Цепочка захвата доказательств

Синхронизированная волна отвечает на следующий вопрос:

> *Были ли наблюдения произведены в сопоставимое системное время и в заранее определенном порядке подачи?*

NOMOS Захват, с другой стороны, задает вопрос:

> *Можем ли мы показать, что отправленный запрос, полученный первый результат, условия продукта и информация о времени были сохранены без изменений?*

Предыдущая архитектура аудита установила, что каждый аудит должен содержать модель или продукт, дату, страну, язык, набор запросов, количество повторений и запись измерений в качестве обязательной логики измерений. Это требование является не только техническим регулированием. Потому что невидимые инструкции, ложные доказательства и контент, представленный как независимый, в конечном итоге могут искажать представление, принимаемое реальными людьми. Этот раздел:

- волна измерений,
- окно синхронизации,
- UTC стандарт времени,
- слоты отправки,
- время завершения и захвата,
- результат первоначального контакта и первый вывод контента,
- правила технического повторного запуска,
- изменения продукта в пределах волны,
- влияние внешних событий и новостей,
- пакет доказательств захвата NOMOS,
- отношение скриншота, необработанного ответа и метаданных,
- проверка времени,
- целостность файлов,
- хэш и цифровая подпись,
- цепочка доказательств,
- конфиденциальность и редактирование,
- публичный манифест доказательств,
- качество и статус валидности волны

определяет. В этой главе пока не рассматривается:

- против какой реальности будут оцениваться атомарные утверждения в ответах ИИ,
- как администраторы будут кодировать ответы,
- пороговые значения перехода и критических ошибок,
- не финализирует окончательные формулы оценки NOMOS

Основной вопрос Глава 12 заключается в:

> *Как мы можем связать тысячи наблюдений реальных пользователей с одним и тем же событием измерения и превратить каждое из них в неизменяемую, проверяемую, защищающую конфиденциальность единицу доказательства?*

NOMOS Вызов

Вы даёте один и тот же запрос тысяче людей. Вы говорите: «Отправьте сегодня в полдень». Один пользователь отправляет его в Стамбуле в 12:03 вечера. Другой пользователь отправляет его в Лондоне в местный полдень. Ещё один видит задание в Токио вечером и выполняет его через восемь часов.

У одного пользователя часы на устройстве были выставлены неправильно. У другого пользователя отключился интернет. Ещё кому-то не понравился ответ, и он нажал кнопку «Сгенерировать заново». Затем они загрузили скриншот самого длинного и самого позитивного ответа. Один пользователь обрезал только середину ответа, запрос не виден. Название ИИ-продукта не видно.

Другой человек отправил тот же скриншот с двумя разными аккаунтами панели. Один пользователь выполнил задание правильно, но загрузил скриншот через два дня. Между тем они обрезали файл и удалили личную информацию. Исходное состояние файла не было сохранено. Другой пользователь в первом ответе ИИ:

Он получил результат «Я не могу ответить на этот вопрос». Исследователь сказал: «Должна была произойти техническая ошибка» и попросил попробовать снова. Второй ответ был правильным, и в отчёт включили только второй ответ. Другой пользователь столкнулся с реальной ошибкой соединения. Запрос был отправлен, но система не сгенерировала контент.

На этот раз исследователь не разрешил повторить попытку. Тот же протокол был применён по-разному к двум пользователям. Теперь, в середине волны измерений, поставщик ИИ:

- поведение веб-поиска,
- система цитирования,
- используемое руководство модели

изменяет все три. Первые 500 человек видят старое состояние системы; следующие 500 человек видят новое. Затем в отчете говорится: «1,000 человек тестировали одну и ту же систему ИИ одновременно». Это утверждение неверно. Пользователей тысяча, но не обязательно существует единая когерентная волна измерений.

Теперь вы применяете хэш ко всем скриншотам. Вы говорите, что это доказывает, что все наблюдения реальны. Что доказывает хэш? Он подтверждает, что файл не был изменен после создания хэша. Один по себе он не доказывает:

- Что изображение получено из реального продукта ИИ
- Что это был первый ответ
- Что оно получено в правильное время
- Что оно не было скопировано у другого пользователя
- Что запрос был правильным

- Что изображение не было отредактировано до создания хэша

Первая предпосылка этого раздела:

> *Измерение одновременно не то же самое, что клик в одну и ту же секунду.*

Второе положение гласит:

> *Скриншот является частью пакета доказательств; он не является всей совокупностью доказательств.*

Её третье положение гласит:

> *Первый неправильный ответ не является ошибкой данных. Если он получен в соответствии с протоколом, это результат измерения.*

Четвертая статья гласит:

> *Хэш усиливает целостность; он сам по себе не создает реальность наблюдения.*

Пятая его статья гласит:

> *Если система существенно меняется за одну волну, тысяча ответов без временной метки не могут считаться принадлежащими одной и той же системе.*

Шестое положение гласит:

> *Цепочка доказательств создаётся не для того, чтобы результат выглядел хорошо, а чтобы позднее можно было повторно проверить, как был получен результат.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — получить наблюдения пользователя GEO-1000 в сопоставимых временных окнах и сохранить каждое наблюдение с полной цифровой цепочкой доказательств. Раздел стандартизирует следующие различия:

- Окно измерения синхронизировано с той же секундой
- Местное время по сравнению со временем UTC
- Волна по сравнению с суб-волной
- Время передачи по сравнению с временем завершения ответа
- Время захвата по сравнению с временем загрузки
- Запланированный слот по сравнению с фактической передачей
- Первый контакт системы по сравнению с первым ответом контента
- Техническая ошибка по сравнению с отклонением системы
- Техническая повторная попытка по сравнению с обновлением ответа
- Лучший выбранный результат с первым подходящим результатом
- Видимость на экране с завершением ответа
- Сырой ответ с экранной копией
- Скриншот против полного пакета доказательств
- Оригинальность наблюдения с целостностью файла
- Отметка времени с хэшем
- Надежный источник времени с местным временем устройства
- Время принятия сервером с загрузкой участника
- Редактированная публичная копия с сырыми доказательствами

- Идентичность участника с идентификатором наблюдения
- Точность ответа с действительностью захвата
- Естественная вариативность вывода с обновлением продукта
- Изменение продукта ИИ с внешним новостным событием
- Прекращение волны с исключением после результата
- Автоматическое измерение бота с захватом пользователя-человека
- API-журнал с записью пользовательский интерфейс
- Телеметрия поставщика с независимым захватом
- Доступ ко всем доказательствам путем их хранения
- Необратимая запись с неизменяемой записью
- Версионированная корректировка неправильного решения с защитой исходных данных

В конце этого раздела каждая проверка GEO-1000 должна быть способна ответить на следующие вопросы:

К какой волне принадлежит это наблюдение?

В какое время был отправлен запрос UTC?

Каков был исходный результат системы?

Когда был завершен ответ и когда он был зафиксирован?

Если был выполнен технический повтор, какова была причина и где была предыдущая попытка?

Показывает ли снимок экрана продукт и полный ответ?

Соответствуют ли друг другу необработанный ответ и визуальное подтверждение?

Когда были сгенерированы хеш и временная метка файлов?

Если доказательство было изменено позже, видна ли история изменений?

Может ли независимый аудитор проверить необработанные доказательства, сохраняя при этом конфиденциальность участников?

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Каждое наблюдение GEO-1000 должно быть связано с зафиксированной измерительной волной до сбора данных; все фактические временные показатели от отправки запроса до первого результата системы, от завершения ответа до захвата и загрузки; условия работы с продуктом ИИ; полный отправленный запрос; первый подходящий результат и все записи целостности должны сохраняться в версионированном пакете доказательств Capture Evidence Package NOMOS. Одно наблюдение GEO-1000:

- снимок экрана
- скопированный текст ответа,
- заявление участника,
- хэш файла

не может считаться действительным сам по себе. Основной пакет доказательств, в той мере, в какой это имеет значение, должен включать:

- ID волны

- Анонимный идентификатор наблюдения участника
- Запись регистра системы ИИ
- Статус панели
- ID и версия запроса
- Назначенный запрос
- Отправленный запрос
- Время отправки
- Первичный контакт с системой
- Время завершения ответа
- Время захвата
- Время загрузки
- Полный необработанный ответ
- Скриншоты полного экрана или последовательные скриншоты
- Справочные и исходные записи
- Журнал технических ошибок и повторных попыток
- Настройки продукта
- Хэши файлов
- Хэш пакета
- Запись проверки времени
- Версия инструмента захвата
- Решение о действительности
- Статус конфиденциальности и редактирования
- Владелец цепочки хранения

3. ЧТО ТАКОЕ ИЗМЕРИТЕЛЬНАЯ ВОЛНА?

Измерительная волна — это комплексное мероприятие по сбору данных, проводимое в рамках заранее определённых:

- ИИ-продуктов,
- целевая популяция,
- версий промптов,
- панелей пользователей,
- временного окна,
- распределения выборки,
- протокола захвата,
- эталонных записей

Волна:

```
Wk=(A,U,P,F,T,C,R,G)
```

может быть представлен следующим образом. Здесь:

- A: примеры продуктов ИИ
- U: целевая популяция пользователей и выборка
- P: набор промптов

- F: состояния панели
- T: структура времени и подачи
- C: протокол захвата
- R: версия реальности для ссылки
- G: управление и ключевая запись

Если любое из этих полей изменится существенно, может потребоваться новая волна или подволна.

4. ТИПЫ ВОЛН

MW-1— ОСНОВНАЯ СИНХРОНИЗИРОВАННАЯ ВОЛНА

Это основное измерение популяции GEO-1000. По умолчанию оно стремится к получению как минимум 1,000 действительных наблюдений Популяционной панели для каждого продукта ИИ.

MW-2— ПОДТВЕРЖДАЮЩАЯ ВОЛНА

Проводится для повторения результата первой волны, для подтверждения конкретной ошибки или для проверки статистической стабильности. Не используется для изменения низкого или высокого балла первой волны. Производит отдельный результат.

MW-3— ИНЦИДЕНТНАЯ ВОЛНА

Проводится после критического или значимого события наблюдателя для целевой и быстрой проверки. Не является глобальной основной волной.

MW-4— ДРЕЙФОВОЕ МОНИТОРИРОВАНИЕ

Отслеживает изменения во времени поведения продукта, объекта, ресурса или языка.

MW-5— ВОЛНА ПОВТОРНОЙ ПРОВЕРКИ КОРРЕКЦИИ

Вырабатывает новые результаты после корректировки или вмешательства. Предыдущая волна не стирается.

MW-6— ВОЛНА С БАЛАНСИРОВАННЫМ МЕСТНЫМ ВРЕМЕНЕМ

Цель — измерить участников в сходных условиях местного времени. Не измеряет единый момент глобальной системы.

MW-7— ВОЛНА КОНТРОЛИРУЕМОЙ ЛАБОРАТОРИИ

Замороженная модель — это контролируемое исследование, проведённое с использованием API или стандартных тестовых аккаунтов. Это не настоящая Панель Пользовательской Популяции.

5. ЧТО ОЗНАЧАЕТ «ОДНО И ТО ЖЕ ВРЕМЯ»?

В GEO-1000 одновременность:

Не означает, что все пользователи кликают в одну и ту же миллисекунду

Функциональное определение:

Наблюдения создаются в пределах предопределённого окна измерения UTC в соответствии с зарегистрированными и независимыми от результатов слотами подачи.

Этот подход учитывает три цели:

- Сохранение статуса продукта близким во времени
- Снижение внезапной искусственной нагрузки на поставщика
- Распределение подач участников воспроизводимым образом

6. СТРУКТУРА ВРЕМЕНИ

Каждая волна должна включать как минимум четыре временных слоя.

6.1. Окно волны

Это диапазон UTC, в пределах которого должны быть сделаны все основные подачи. Пример:

12:00–13:00 UTC

6.2. микрослот

Это более короткий временной интервал, отведённый участнику для отправки запроса. Пример:

12:15–12:20 UTC

6.3. Дополнительное время на выполнение

Если запрос отправлен в пределах слота, заранее установленное время для завершения ответа составляет.

6.4. Время захвата и загрузки

Время, разрешенное для создания доказательства и загрузки его в защищённую систему после завершения ответа.

7. ОСНОВНОЕ ПРАВИЛО СИНХРОНИЗИРОВАННОЙ ВОЛНЫ КАНДИДАТА

ПРАВИЛО КАНДИДАТА — ОСНОВНОЕ ПРАВИЛО КАНДИДАТА

Для общих потребительских чатов можно использовать следующую структуру в качестве основной синхронизированной волны кандидата:

- Окно подачи: 60 минут
- Микрослоты: 12× 5 минут
- Отправка участника: в рамках назначенного им микрослот
- Дополнительное время для завершения ответа: до 20 минут после подачи
- Генерация хеша захвата: предпочтительно в течение 5 минут после завершения ответа
- Загрузка доказательств: не позднее чем через 60 минут после захвата
- Сообщение о технических проблемах: в пределах той же волны

Эти длительности являются универсальными и не окончательными. Они могут изменяться в зависимости от следующих условий:

- Время отклика продукта AI
- Мобильное соединение
- Инфраструктура страны
- Формат длинного ответа
- Требование доступности
- Лимит использования продукта

Время, которое используется, должно быть зафиксировано до сбора данных.

8. ПОЧЕМУ МИКРО-СЛОТЫ?

Одновременная отправка тысячами участников:

- Необычная нагрузка на продукт AI,
- ограничение скорости,
- очередь,
- техническая ошибка,
- увеличение времени отклика

может быть создано. В этом случае аудит не может измерить нормальное поведение продукта, но:

может измерить искусственное событие нагрузки, созданное аудитом

Микрослоты:

- распределяют нагрузку,
- сохраняют одинаковое общее системное время,
- делают порядок подачи независимым от результата,

позволяют моделировать эффект времени.

9. НАЗНАЧЕНИЕ СЛОТОВ

Участники должны быть назначены в слоты:

- страны,
- языка,
- Продукт ИИ,
- план,
- сбалансированным и предварительно случайным образом по:

статус панели. Один микрослот не должен заполняться:

- для конкретной страны,
- для конкретного языка,
- для конкретного продукта

пользователями. В противном случае эффект времени будет смешан с эффектом группы.

10. СПРАВЕДЛИВОСТЬ ЛОКАЛЬНОГО ВРЕМЕНИ

Окно UTC может быть для некоторых участников:

- ночью,
- рабочие часы,
- время молитвы или отдыха,
- труднодоступные часы

На эта ситуация может влиять:

- уровень участия,
- состояние соединения,
- внимание пользователя,
- тип устройства

Можно использовать два дополнительных метода.

10.1. Вращающиеся глобальные волны

Основные волны проводятся в разное время UTC. Пример:

- Волна 1: 12:00 UTC
- Волна 2: 20:00 UTC
- Волна 3: 04:00 UTC

Таким образом, один и тот же регион не несет одинаковую нагрузку по местному времени в каждой волне.

10.2. Подволна с балансировкой по местному времени

Участники измеряются в аналогичном временном диапазоне по их местному времени. Это исследование:

- усталость пользователя,
- имеет ценность для

ежедневного поведения пользователей.

Это не результат единого глобального момента системы.

11. АРХИТЕКТУРА ТРИВОЛНОВОЙ СТАБИЛЬНОСТИ

АРХИТЕКТУРА КАНДИДАТА — АРХИТЕКТУРА КАНДИДАТА

Можно предложить как минимум три основные волны для полной реализации GEO-1000:

W1 — Начальная волна

Генерирует начальную оценку популяции.

W2 — Краткосрочное повторение

Тестирует краткосрочную стабильность продукта и измерения. Интервал кандидата: 24–72 часов после W1

W3 — Среднесрочная стабильность

Измеряет, сохраняется ли представление о продукте и субъекте в долгосрочной перспективе. Предполагаемый интервал: 14–30 дней после W1. Конкретные интервалы должны быть определены до сбора данных. Если он будет изменен из-за обновления продукта или события, требуется запись новой версии.

12. 30,000 НАБЛЮДАЕМЫЙ ОСНОВАТЕЛЬ ДИЗАЙН

Для продуктов ИИ 1,000 основные наблюдения на продукт и для трех основных волн:

```
10×1,000×3=30,000
```

Целью являются действительные основные наблюдения Панели Населения. Дополнительно:

- Наблюдатель страны
- Языковая справедливость
- Доступность
- Инцидент
- Повторное тестирование

панели могут быть добавлены поверх этого числа. 30,000:

- не является числом приглашений,
- не является числом файлов,

не количество скриншотов. Это действительная основная цель наблюдения за продуктом.

13. ВЕКТОР ВРЕМЕНИ НАБЛЮДЕНИЯ

Для каждого наблюдения вектор времени может быть определён как кандидат следующим образом:

```
T_i= (t_i^assign, t_i^open, t_i^submit, t_i^first, t_i^complete, t_i^capture, t_i^upload,
t_i^validate)
```

Здесь:

- tiassign: время назначения задачи пользователю
- tiopen: время открытия задачи
- tisubmit: время отправки запроса
- tifirst: время начала появления первого видимого результата системы
- ticomplete: время завершения ответа
- ticapture: время фиксации доказательства
- tiupload: время, когда пакет достигает защищённой системы
- tivalidate: время проверки действительности фиксации

Не все эти времена могут быть непосредственно наблюдаемы на каждом пользовательском интерфейсе. Неизвестные поля должны оставаться НЕИЗВЕСТНЫМИ.

14. ОСНОВНЫЕ ВРЕМЕННЫЕ МЕТРИКИ

Задержка отклика:

```
L_i= t_i^complete- t_i^submit
```

Задержка захвата:

```
C_i= t_i^capture- t_i^complete
```

Задержка загрузки:

```
U_i= t_i^upload- t_i^capture
```

Отклонение слота:

```
D_i= t_i^submit- t_i^{slot-midpoint}
```

можно рассчитать следующим образом. Эти метрики:

- нагрузка системы,
- задержка участника,
- уверенность в доказательстве,
- целостность волны

с точки зрения оценки.

15. НЕ МОЖЕТ БЫТЬ УПРАВЛЯЕМО СОГЛАСНО ЗАДЕРЖКЕ

Ответ, выполненный за длительный период:

- не может быть вынуждено к категориям

  истинно,
- ложно,
- длинным,
- может быть процитирован.

Из-за задержки правило исключения:

- до сбора данных,
- независимо от содержания ответа

должно быть определено. Следующее поведение запрещено: исключение отрицательного ответа из-за его позднего выполнения; защита положительного ответа, если он выполнен с той же задержкой.

16. СОСТОЯНИЯ ВОЛНЫ

WS-0— ЗАПЛАНИРОВАНО

Волна запланирована. Она еще не зафиксирована.

WS-1— ЗАФИКСИРОВАНО

Образец, промпт, продукт, время и правила захвата зафиксированы.

WS-2— ОТКРЫТО

Окно подачи началось.

WS-3— ЗАКРЫТО

Прием новых подач завершен. Время выполнения и загрузки может продолжаться.

WS-4— ЗАВЕРШЕНО

Проверки на поле и первоначальный захват выполнены.

WS-5— РАЗДЕЛЕНО

Он был разделен на подволны из-за изменений в материальной системе или протоколе.

WS-6— ПРИОСТАНОВЛЕНО

Временно приостановлено из-за проблем с безопасностью, продуктом или инфраструктурой.

WS-7— ПРЕРВАНО

Прекращено до завершения волны. Результаты могут быть ограниченными или непригодными.

WS-8— АННУЛИРОВАНО

Недействительно для основного прогноза из-за ранее определенного серьезного дефекта протокола. Первичные наблюдения не удаляются.

WS-9— В АРХИВЕ

Версия закрыта, а цепочка доказательств архивирована.

17. КЛАССЫ СОСТОЯНИЯ ВОЛНЫ

WH-0— НЕ ОЦЕНИВАЛАСЬ

Состояние волны не было оценено.

WH-1— СОМНЕТЕЛЬНОЕ СОСТОЯНИЕ

Произошли изменения в материальной системе, время или ошибка захвата. Может быть непригодно для основного результата.

WH-2— ЧАСТИЧНО

Некоторые слои или слоты недостаточны. Может быть получен уменьшенный результат.

WH-3— ПРИЕМЛЕМО

Основное время, система и условия захвата Adequately сохранены.

WH-4— СИЛЬНО

Наблюдаются высокая уверенность захвата, сбалансированное покрытие слотов и низкие потери протокола.

WH-5— ВОСПРОИЗВЕДЕНО

Аналогичное состояние волны повторялось несколько раз на независимых панелях. Состояние волны не зависит от того, высокий или низкий балл ИИ.

18. ИЗМЕНЕНИЯ ПРОДУКТА ВНУТРИ ВОЛНЫ

Продукт с ИИ может существенно изменяться в пределах окна измерения. Сигналы:

- Изменение марки модели
- Изменение интерфейса
- Активация или деактивация веб-функции
- Изменение формата цитирования
- Внезапный сбой в структуре ответов
- Объявление поставщика
- Различное поведение системы одновременно у многих пользователей
- Прерывание или перезапуск продукта

В этом случае: время изменения оценивается. Сравниваются конфигурационные отпечатки системы. Волна WS-5 может быть переведена в статус SPLIT. Предыдущие и последующие наблюдения будут отдельными подволнами. Если нужно получить единый показатель, статус системы моделируется как отдельный фактор. Существенная неопределённость документируется в публичной записи.

19. ЗАПИСЬ О ВНЕШНЕМ СОБЫТИИ

Мир может измениться, даже если продукт ИИ не меняется. Во время измерения:

- приобретение компании,
- запуск продукта,
- значимые новости,
- юридическое решение,
- кризис,
- финансовое раскрытие,
- вирусный контент

может произойти. Это событие:

- веб-ресурсы,
- результаты извлечения,
- комментарий к запросу пользователя,
- ответ ИИ

может быть изменено. Каждая волна, в той мере, в какой это релевантно:

Регистр внешних событий

должны содержать.

20. КАК УПРАВЛЯЮТСЯ РЕЗУЛЬТАТЫ ВНЕШНИХ СОБЫТИЙ?

Внешнее событие:

- если известно до волны, обновляется пакет ссылок,
- если происходит во время волны, отмечается с отметкой времени,

Если замечено после волны, добавляется ретроспективная запись о событии. Следующее поведение запрещено: исключение наблюдений из-за новостей, которые снижают оценку; сохранение их, когда это новости, повышающие оценку. Сами события могут быть частью состояния системы в реальном мире. При необходимости проводится анализ с учетом подволны или корректировки события.

21. РЕЗУЛЬТАТ ПЕРВОГО КОНТАКТА

Первый результат системы, с которым сталкивается пользователь после отправки запроса:

Результат первого контакта

называется. Результат первого контакта:

- Полный ответ
- Отказ
- Техническая ошибка
- Лимит использования
- Прерывание соединения
- Бесконечная загрузка
- Уточняющий вопрос
- Нет результата

возможен. Это событие является частью реального пользовательского опыта. Оно не может быть удалено без уведомления.

22. ПЕРВЫЙ ДОПУСТИМЫЙ ВЫХОД КОНТЕНТА

Первый материальный ответ, сгенерированный при ранее разрешенной попытке после технической ошибки:

Может называться Первым Допустимым Выходом Контента

Этот выход: не заменяет результат первоначального контакта, он записывается вместе с ним. Пример:

- Первичный контакт: ограничение скорости
- Пять минут спустя, ранее разрешённая техническая повторная попытка
- Первый ответ содержимого

Отчет должен показывать следующее вместе:

- Первый пользовательский опыт: ограничение скорости
- Первое представление содержимого: указанная реакция

23. ПЕРВОЕ ПОДХОДЯЩЕЕ ПРАВИЛО ВЫВОДА

Если продукт ИИ дает существенный ответ на первый запрос:

- даже если он неверный,
- даже если он короткий,
- даже если он включает отказ,

- даже если не предоставляет ссылки,
- даже если неправильно идентифицирует бренд

воспроизведение не может быть сделано. Основное наблюдение — первый вывод. Пользователь:

- не может выбрать новый ответ с помощью:
- регенерации,
- переписывания,
- повторной попытки,
- другого ответа,

продолжения

24. ТЕХНИЧЕСКАЯ ПОВТОРНАЯ ПОПЫТКА

Техническая повторная попытка может использоваться только при ранее определённых технических условиях. Допустимые причины кандидата:

- Прерывание соединения до начала ответа
- Код ошибки сервера
- Интерфейс не отправляет запрос
- Сбой приложения
- Подтверждённый временный сбой системы

Не может быть причиной технической повторной попытки:

- Неправильный ответ
- Отказ
- Краткий ответ
- Отсутствие ссылки
- Бренд не упомянут
- Критическая ошибка
- Аудированная организация не довольна ответом

25. ПОВТОРНАЯ ПОПЫТКА ПРИ ЧАСТИЧНОМ ОТВЕТЕ

Если система создала часть содержимого материала и затем остановилась:

- частичный вывод сохраняется,
- сохраняется как TRUNCATED_RESPONSE

если необходимо повторить попытку, создается новая связанная попытка. Ее нельзя удалить, как если бы частичного ответа не было.

26. ЦЕПОЧКА ПОПЫТОК

Назначение наблюдения может включать несколько технических попыток:

```
A_i→ (Attempt_{i,1}, Attempt_{i,2}, …)
```

Каждая попытка:

- время,
- содержит состояние продукта,
- хэш запроса,
- результат системы,
- причину технической ошибки

Не перезаписывается другой попыткой.

27. ЧТО ТАКОЕ NOMOS CAPTURE?

NOMOS Capture — это открытый стандарт данных и процедур, который сохраняет целостность наблюдения от предоставленного пользователю запроса до архивирования пакета доказательств. NOMOS Capture:

- это не просто приложение для скриншотов,
- не обязательно быть проприетарным программным обеспечением, принадлежащим только NobleJackal,

не зависит только от одного расширения браузера. Совместимое приложение NOMOS Capture:

- использует открытую схему данных,
- соблюдает правила времени и целостности,
- уважает границы конфиденциальности,
- сохраняет цепочку доказательств

должны быть выполнены. Независимые исследователи должны иметь возможность разрабатывать свои собственные совместимые инструменты. Если стандарт может быть реализован только с программным обеспечением основателя, независимое воспроизведение будет ослаблено.

28. ЧЕТЫРЕ ФУНКЦИИ NOMOS CAPTURE

28.1. Выполнить задачу

Правильно:

- пользователь,
- запрос,
- Продукт ИИ,
- слот

обеспечивает соответствие.

28.2. Запечатлеть наблюдение

Логирует запрос, начальный результат системы, ответ и соответствующий пользовательский интерфейс.

28.3. Поддерживать целостность

Генерирует хеш файла, отметку времени и связь пакетов.

28.4. Перенос доказательств

Связывает необработанные, ограниченные и общедоступные уровни доказательств.

29. МЕТОДЫ ЗАХВАТА

CM-1— РУКОВОДСТВО ПО РУЧНОМУ ЗАХВАТУ

Участник предоставляет скриншот и текст необработанного ответа согласно инструкциям. Это имеет низкую техническую нагрузку. Подвержено человеческой ошибке.

CM-2— ПРИЛОЖЕНИЕ ДЛЯ ПОМОЩИ ЗАХВАТА

Приложение:

- копирование запросов,
- ведение учета времени,
- загрузка файлов,
- генерация хешей

функции поддерживаются. Не мешает продукту ИИ.

CM-3— ИНСТРУМЕНТАЛЬНОЕ ЗАХВАТ ИНТЕРФЕЙСЫ ПОЛЬЗОВАТЕЛЯ

Интеграция с браузером или приложением:

- отправленная промпт,
- метаданные продукта,
- текст ответа,
- времена

захватываются напрямую. Обеспечивает высокую целостность. Следует оценить условия использования и политику конфиденциальности.

CM-4— ЭКСПОРТ ИЛИ ТЕЛЕМЕТРИЯ ПОСТАВЩИКА

Используется экспорт сеанса провайдера или проверенная телеметрия. Может сильно поддерживать независимый захват. Сама по себе может не обеспечивать независимость.

CM-5— КОНТРОЛИРУЕМЫЙ API ИЛИ ЗАХВАТ В ЛАБОРАТОРИИ

Используются журналы API или замороженной системы. Это не наблюдение за живым пользовательский интерфейс. Относится к отдельной лабораторной панели.

30. ПРЕДЕЛ АВТОМАТИЗАЦИИ

Автоматизация:

- отправка промптов,
- ведение учета времени,
- генерация хэша,
- целостность файла

можно использовать для. Вместо участника:

- аккаунт бота,

- автоматизированный массовый запрос,
- непользовательская сессия

генерация не создает настоящую Панель Населения человека. Автоматический аудит системы осуществляется отдельно:

CONTROLLED_AUTOMATION_PANEL

должен быть записан как.

31. ЗАХВАТ УРОВНЕЙ ДОВЕРИЯ

NCL-0— НЕТ ПОДТВЕРЖДАЕМОГО ЗАХВАТА

Есть только заявление участника. Не может быть включено в основной результат GEO-1000.

NCL-1— ЧАСТИЧНЫЕ ДОКАЗАТЕЛЬСТВА

Существует текст ответа или ограниченный скриншот. промпт или состояние системы не могут быть полностью проверены.

NCL-2— ДОКАЗАНО С ЭКРАНА

Запрос, ответ и главный интерфейс продукта проверены с помощью скриншота. Метаданные и уверенность во времени могут быть ограничены.

NCL-3— ПОЛНЫЙ ПАКЕТ ДОКАЗАТЕЛЬСТВ

Полный ответ, промпт, запись продукта, время, скриншот и хэш доступны. Минимальный уровень для основного анализа — кандидат.

NCL-4— ИНСТРУМЕНТАЛЬНО И ПОДПИСАННО ЗАХВАЧЕННО

Инструмент захвата зарегистрировал промпт, ответ, время и метаданные напрямую; пакет цифрово подписан.

NCL-5— НЕЗАВИСИМО ПРОВЕРЕННЫЙ МНОГОИСТОЧНИКОВЫЙ ЗАХВАТ

Пакет захвата:

- независимая проверка,
- экспорт поставщика,
- второй источник доказательств

поддерживается. Основные GEO-1000 реальные наблюдения пользователя как минимум:

NCL-3

должен нацеливаться на этот уровень.

32. NOMOS ПАКЕТ ДОКАЗАТЕЛЬСТВ ЗАХВАТА

Для каждого наблюдения пакет содержит следующие компоненты в качестве кандидатов:

32.1. Манифест наблюдения

Включает все поля идентификации и статуса наблюдения.

32.2. Артефакт промпта

Полный текст промпта, присвоенной и отправленной, версия и ее хэш.

32.3. Артефакт необработанного ответа

Полный необработанный ответ, созданный продуктом ИИ.

32.4. Артефакт визуальных доказательств

Показывает изображения или последовательные изображения запроса, ответа и интерфейс продукта.

32.5. Артефакт состояния системы

Показывает план, интерфейс, метку модели, память, веб и другие материальные состояния системы.

32.6. Артефакт временных доказательств

Содержит время отправки, завершения, захвата и загрузки.

32.7. Журнал попыток

Показывает цепочку технических ошибок и повторных попыток.

32.8. Манифест целостности

Содержит хэши и размеры всех файлов.

32.9. Манифест конфиденциальности

Показывает статус редактирования, доступа и хранения.

32.10. Запись проверки

Указывает, кто и когда оценивал достоверность захвата.

33. МИНИМАЛЬНОЕ СОДЕРЖАНИЕ СКРИНШОТА

В пределах соответствия визуальные доказательства должны показывать:

- Продукт ИИ или идентичность интерфейсы
- Полный запрос
- Начало ответа
- Конец ответа
- Справочные поля
- Сообщение об отклонении или ошибке
- Видимая метка модели
- Доказательства того, что сессия новая или находится в соответствующем состоянии
- Настройка материальной системы
- Состояние интерфейса, указывающее, что ответ не был обрезан

34. ДОЛГИЕ ОТВЕТЫ

Если ответ не помещается на одном экране:

- полный скриншот страницы,
- скриншот с возможностью прокрутки,
- перекрывающиеся последовательные скриншоты,
- запись экрана,
- прямой экспорт текста

можно использовать. Последовательные визуальные элементы:

- не должны оставлять пробелов,
- должны перекрываться друг с другом,

должны иметь номера последовательности. Только положительная или соответствующая секция не может быть обрезана.

35. ЗАПИСЬ ЭКРАНА

Запись экрана:

- отправка промптов,
- потоковый ответ,
- статус завершения,
- обновление не выполнено

может быть показано. Однако:

- более личная информация,
- уведомление,
- частное приложение,
- ID аккаунта

может быть захвачено. Поэтому запись экрана:

- может не быть обязательной по умолчанию,
- может использоваться в подвыборке с высоким риском или для проверки,

требует подготовки к конфиденциальности перед записью.

36. ИСХОДНЫЙ ТЕКСТ ОТВЕТА

В дополнение к скриншоту полный текст ответа должен храниться отдельно. Порядок предпочтения:

- Официальный экспорт или функция копирования продукта
- Прямой захват текста с помощью инструмента захвата
- Доступность дерева или извлечение на основе DOM
- Текст, скопированный пользователем
- OCR и человеческая проверка

OCR должен быть только последним вариантом. Если существует существенная разница между автоматическим извлечением и визуальной проверкой, требуется ревизия.

37. ДИНАМИЧЕСКИЕ ОТВЕТЫ

Некоторые продукты, после завершения ответа:

- могут добавить цитату,
- могут загрузить карточку источника,
- могут переформатировать текст,

могут добавить предупреждение о безопасности. Можно использовать два момента захвата:

- Первое завершение захвата
- Захват после краткой стабилизации рендера

Если есть существенная разница, обе версии сохраняются. Первоначальное взаимодействие пользователя не удаляется.

38. КАК ПОНЯТЬ, ЧТО ОТВЕТ ЗАВЕРШЕН?

Индикаторы завершения кандидата:

- Потоковая передача останавливается
- Индикатор загрузки закрывается
- Появляется кнопка «Перегенерировать» или эквивалентная
- Сигнал завершения продукта
- Предустановленный период тишины
- Запись завершения через API

Правила завершения на основе продукта должны быть найдены в Регистре систем ИИ.

39. ЕСЛИ ПОЛЬЗОВАТЕЛЬ ПРЕКРАТИЛ ОТВЕТ

Пользователь мог случайно или намеренно:

- остановить,
- отменить,
- прекратить приложение

совершить действие. Эта ситуация фиксируется как:

ПОЛЬЗОВАТЕЛЬ_ПРЕРВАЛ

СЛУЧАЙНОЕ_ПРЕРЫВАНИЕ

НЕИЗВЕСТНЫЙ ПЕРЕРЫВ

Ответ, остановленный пользователем, не может считаться полным системным результатом. Правило замены участника или повторная попытка должны быть определены заранее.

40. промпт И ОТВЕТ НАХОДЯТСЯ В ОТДЕЛЬНЫХ ФАЙЛАХ

Если подска□ зка и ответ находятся в отдельных изображениях:

- они должны быть связаны с одной сессией,
- с одним и тем же идентификатором наблюдения,
- и с непрерывной цепочкой времени

В противном случае может быть невозможно подтвердить, что запрос и ответ действительно принадлежат одной и той же беседе.

41. ЦЕЛОСТНОСТЬ ФАЙЛА

Для каждого файла должен быть сгенерирован криптографический хеш. Кандидат [K25]:

```
h_j= H(file_j)
```

Хеш пакета:

```
h_bundle= H(JCS(manifest(file_id, порядок, размер, алгоритм, h_j))) [K14; K25]
```

может быть создан. Здесь H — это опубликованный и признанный безопасным хеш-алгоритм. Алгоритм:

- версионированная,
- открытая,
- может быть изменён в будущем при необходимости.

должен быть.

42. ЧТО ДОКАЗЫВАЕТ ХЕШ?

Если записанное значение хэша хранится надежно, повторное хэширование того же файла с использованием того же алгоритма помогает обнаружить изменения в последовательности байтов [K25]. Это сравнение подтверждает утверждение о том, что файл сохранил ту же последовательность байтов после генерации эталонного хэша; оно не доказывает источник файла, что он был создан в правильное время, или его подлинность до хэширования. Один только хэш не доказывает:

- Что файл поступил от настоящего продукта ИИ
- Что он принадлежит правильному пользователю
- Что он был создан в правильное время
- Не был организован до хэширования
- Что это первый результат
- Что запрос был правильным

Следовательно, хэш:

> *Является доказательством целостности.*

Сам по себе:

> *Не является доказательством оригинальности.*

43. ЦИФРОВАЯ ПОДПИСЬ

Инструмент захвата или уполномоченный проверяющий:

- пакет наблюдения,

- запись времени,
- версия транспортного средства

может подписывать цифровым образом. Подпись:

- каким транспортным средством или учреждением был создан пакет,
- что он не изменился после подписи

усиливает. Держатель подписи мог подписать неверные или поддельные данные. Следовательно, сама по себе подпись не гарантирует подлинность.

44. НАДЕЖНОЕ ВРЕМЯ [K26]

Локальные часы устройства:

- неправильно настроены,
- изменены вручную,
- не синхронизированы

могут возникнуть. Надежная отметка времени должна содержать несколько сигналов:

- Время приёма сервером захвата
- Надежный сервис времени UTC
- Время устройства и запись отклонений
- Видимое время в продукте ИИ, если таковое имеется
- Время создания файла
- Цифровая временная метка

45. ОТКЛОНЕНИЕ ВРЕМЕНИ

Разница между временем устройства участника и надежным временем UTC:

```
O_i^clock= t_i^device- t_i^reference
```

может быть записана как. Если есть существенное отклонение:

- время сервера используется в качестве основного ориентира,
- время устройства сохраняется как исправленная запись,

неопределенность отмечается. Участника не следует просить вручную изменять время на устройстве.

46. ВРЕМЯ ЗАГРУЗКИ НЕ ЯВЛЯЕТСЯ ВРЕМЕНЕМ СОЗДАНИЯ

Загрузка файла на сервер в 14:00 не доказывает, что скриншот был сделан в 14:00. Поэтому:

- время создания,
- время генерации хеша,
- время загрузки

должно регистрироваться отдельно. Длительные задержки при загрузке могут снизить доверие к захвату. Запись не обязательно должна быть автоматически недействительной.

47. ЗАДЕРЖКА ЗАГРУЗКИ

Из-за проблем с подключением пользователь может не иметь возможности сразу загрузить доказательства. Если поздняя загрузка допускается:

- локальный хэш на момент захвата,
- время доверенного устройства,
- офлайн-подписанная запись,
- причина задержки загрузки

могут быть необходимы. Только заявление пользователя: «Я сделал это изображение вчера» недостаточно для основного уровня NCL-3.

48. ЦЕПОЧКА ДОКАЗАТЕЛЬСТВ

Цепочка доказательств наблюдения в качестве кандидата состоит из следующих этапов:

- Назначение задачи
- Фиксация промпта
- Соответствие пользователя и продукта
- Открытие слота
- Отправка промпта
- Результат первичного контакта
- Завершение ответа
- Захват
- Локальная запись целостности
- Безопасная загрузка
- Время принятия сервером
- Хэш пакета
- Проверка протокола
- Редактирование
- Арбитраж
- Архив
- Публичный манифест

Каждый этап должен быть связан с предыдущим этапом.

49. ПРЕРЫВАНИЕ ЦЕПОЧКИ

Цепочка доказательств может быть прервана в следующих случаях:

- Отсутствие записи запроса
- Начальный вывод не может быть проверен
- Ответ не соответствует визуальному отображению
- Время захвата неизвестно
- Файл был изменён позже
- Дублирующийся идентификатор участника

- Продукт и план неизвестны
- Наблюдение вне волны
- Хэш не соответствует файлу
- Редактирование перезаписало исходный файл
- Неизвестно, кто это проверил

Тип нарушения определяет статус действительности.

50. ИСХОДНЫЕ ДОКАЗАТЕЛЬСТВА НЕ МОГУТ БЫТЬ ИЗМЕНЕНЫ; РЕШЕНИЕ МОЖЕТ ИЗМЕНИТЬСЯ

Пакет исходных доказательств:

- не должен быть перезаписан,
- не должен быть тихо исправлен,

не должен быть удален и заменен новым. Однако, касательно наблюдения:

- действительность,
- принятие решений,
- классификация,
- соответствие

решение может измениться. Новое решение:

- новая версия,
- обоснование,
- дата,
- принимающий решение

должно нести. Это различие является фундаментальным правилом управления:

> *Доказательства могут оставаться неизменными; интерпретация может развиваться вместе с доказательствами.*

51. РЕДАКЦИЯ

Следующие поля могут быть скрыты в публичной или арбитражной копии:

- Имя и фамилия
- Электронная почта
- Изображение профиля
- Идентификатор аккаунта
- Личные переписки
- Уведомления
- Корпоративное поле
- Название конфиденциального файла
- Точное местоположение

Редакция:

- не должна проводиться на исходном файле,

- должна производить новый производный файл,
- должна иметь собственный хеш,

должна фиксировать, какое поле было скрыто и почему.

52. ТРИ УРОВНЯ ДОКАЗАТЕЛЬСТВ

52.1. Исходные ограниченные доказательства

Это полностью сырые доказательства. Доступ к ним имеют только уполномоченные аудиторы и необходимые исследовательские роли.

52.2. Слой аудиторских доказательств

Содержит доказательства, необходимые для независимых арбитров и команд воспроизведения. Личные данные сведены к минимуму.

52.3. Манифест общественных доказательств

Для публики:

- ID наблюдателя,
- страна и язык,
- Продукт ИИ,
- время,
- хеш доказательства,
- статус действительности,
- статус редактирования

представляются. Сырой личный контент не обязательно публикуется.

53. ПУБЛИЧНЫЙ МАНИФЕСТ ДОКАЗАТЕЛЬСТВ

Для каждой волны публичный манифест может включать следующие поля:

- ID волны
- Версия протокола
- Количество продуктов
- Целевая и фактическая численность наблюдений
- Охват стран и языков
- Покрытие слотов
- Уровни уверенности захвата
- Технические ошибки и уровень отклонений
- Число недействительных наблюдений
- Хэши пакетов наблюдений
- Корень целостности волны
- События изменения
- Примеры, опубликованные с цензурой
- Процедура независимого доступа
- Статус хранения доказательств

54. КОРЕНЬ ЦЕЛОСТНОСТИ ВОЛНЫ

КАНДИДАТСКАЯ КОНЦЕПЦИЯ

Пусть хэш каждого пакета наблюдений будет hi. Из них корень волны с деревом Меркла или эквивалентной структурой:

RW

может быть создан. Корень волны может быть опубликован после закрытия измерения. Этот метод:

- делает позднее тихое изменение списка наблюдений в волне,
- добавление нового наблюдения,
- изменение существующего наблюдения

более заметными. Корень Меркла:

- сам по себе не доказывает, что наблюдения реальны,
- или что решение арбитража правильное

Это укрепляет целостность списка.

55. ДОБАВЛЕНИЕ И УДАЛЕНИЕ НАБЛЮДЕНИЙ

После закрытия волны:

- добавление нового наблюдения,
- удаление существующего наблюдения,
- изменение файла

требует новой версии манифеста. Старый корень целостности сохраняется. Для удаляемого наблюдения:

- причина,
- дата,
- лицо, принимающее решения
- влияние на старую и новую оценку

записываются.

56. ЗАПРОС НА ИСЧЕЗНОВЕНИЕ УЧАСТНИКА

Этика исследований или рамки защиты данных могут позволить участнику удалить свои данные. В этом случае:

- сырые данные могут быть безопасно удалены или сделаны недоступными,
- пустая запись WITHDRAWN может быть сохранена в публичном манифесте,
- оценка волны может быть пересчитана с новой версией

старый результат и причина изменения сохраняются. Неизменяемость не означает устранение прав человека.

57. СТАТУСЫ ДОПУСТИМОСТИ ЗАПИСИ НАБЛЮДЕНИЯ

CV-0— НЕ ПРОВЕРЕНО

Запись ещё не была оценена.

CV-1— ДЕЙСТВИТЕЛЬНО

Соответствует всем обязательным условиям записи и времени.

CV-2— УСЛОВНО ДЕЙСТВИТЕЛЬНО

Имеется ограниченное несоответствие. Указано, в каких анализах она может быть использована.

CV-3— ЧАСТИЧНОЕ ДОКАЗАТЕЛЬСТВО

Наблюдение может использоваться как событие. Оно может не входить в основной расчёт населения.

CV-4— ВНЕ ВОЛНЫ

промпт была отправлена за пределами допустимой временной структуры.

CV-5— НЕСООТВЕТСТВИЕ промпта ИЛИ РЕЗУЛЬТАТА

Существует существенное несоответствие между назначенной промптом, поданной промптом или зафиксированным ответом.

CV-6— ДУБЛИКАТ

То же самое наблюдение или запись пользователя было повторено.

CV-7— ПОДОЗРЕНИЕ В МАНИПУЛЯЦИИ

Существует нерешённое сомнение относительно целостности файла или метаданных.

CV-8— СФАБРИКОВАНО ИЛИ ИЗМЕНЕНО

Наблюдение не является реальным пользовательским опытом или было намеренно изменено.

CV-9— ОТЗЫВ

Оно было исключено из анализа из-за процессов, связанных с участником или этических процессов.

58. ПРАВИЛЬНОСТЬ ЗАПИСИ ДОЛЖНА БЫТЬ ОТДЕЛЕНА ОТ СЕМАНТИЧЕСКОЙ ТОЧНОСТИ

Наблюдение:

- полностью верное с точки зрения записи,
- сильно неверное с точки зрения содержания

могло быть. Другое наблюдение:

- правильный ответ,
- может содержать недостающие или поддельные данные

Следовательно, есть два отдельных решения:

- Достоверность захвата
- Семантическое решение

Валидатор захвата должен знать как можно меньше о балле успешности ответа.

59. РОЛЬ ВАЛИДАТОРА ЗАХВАТА

Валидатор захвата проверяет:

- Соответствие промпта
- Волна измерений и временной интервал
- Идентичность продукта
- Правило первого вывода
- Целостность экрана
- Согласованность исходного текста
- Техническая повторная попытка
- Хэш и время
- Конфиденциальность

Не принимает решения по следующим вопросам: была ли компания описана правильно? Подтверждает ли цитата утверждение? Есть ли критическая ошибка? Проходит ли ответ проверку? Это ответственность следующего уровня проверки.

60. ПОВТОР СКРИНА ЭКРАНА

Та же визуальная информация:

- случайно загружена дважды,
- использована несколькими пользователями,
- повторяется среди поставщиков панелей

может возникнуть. Проверки:

- Хэш файла
- Визуальное сходство
- Исходный текст ответа
- Метаданные времени и сессии
- Уникальность пользователя

можно выполнить. Один и тот же продукт ИИ может дать точно такой же ответ дословно некоторым пользователям. Только один и тот же текст не является доказательством мошенничества. Один и тот же файл или один и тот же визуальный след является более сильным сигналом.

61. СХОДСТВО ФАЙЛОВ И ЛОЖНО-ПОЛОЖИТЕЛЬНЫЙ РЕЗУЛЬТАТ

Та же аппаратура и интерфейс:

- похожие визуальные элементы,
- тот же размер экрана,
- та же компрессия

может производить. Алгоритм визуального сходства не должен выполнять автоматическое исключение. Требуются несколько сигналов и проверка человеком.

62. ТЕЛЕМЕТРИЯ ПРОВАЙДЕРА

Поставщик ИИ:

- время сессии,
- модель,
- запрос,
- ответ,
- использование инструмента

может предоставлять надежные журналы об этом. Эта запись усиливает независимое фиксирование. Однако провайдер:

- измеряемая сторона,
- владелец системы,
- выгодоприобретатель

может быть. Телеметрия провайдера: должна использоваться вместе с независимыми доказательствами пользователя, а не заменять их.

63. ЕСЛИ ДАННЫХ ОТ ПРОВАЙДЕРА НЕТ

Доступ к журналам поставщика может быть недоступен в большинстве работающих потребительских продуктов. Это не делает GEO-1000 невозможным. Можно использовать следующий состав:

- Инструментальный захват пользователя
- Отметка времени сервера
- Доказательство полного экрана
- Необработанный ответ
- Системные настройки
- Хэш и подпись
- Подвыборка для независимой проверки

Уровень доверия к доказательству указан соответствующим образом.

64. ЗАХВАТ В Панель пользователей в естественных условиях

Инструмент захвата в естественной панели:

- особое содержание инструкции,
- предыдущие разговоры,
- идентификация учетной записи

не следует захватывать без необходимости. Предпочтительный процесс: пользователь открывает новый чат естественного аккаунта. Capture фиксирует только окно задачи. Статус персонализации берется отдельно как высокоуровневые метаданные. Личная история остается вне захвата. Публичная копия редактируется.

65. ЗАХВАТ СОСТОЯНИЯ НАТУРАЛЬНОГО ОБЩЕНИЯ

Если контекст текущего разговора измеряется:

- только необходимый диапазон предыдущих ходов,
- явное согласие,
- строгое редактирование,
- ограничение доступа

требуются. Вся личная история чата по умолчанию не может быть собрана.

66. МОБИЛЬНЫЙ ЗАХВАТ

На мобильных устройствах:

- полноэкранный захват,
- скриншот с возможностью прокрутки,
- версия приложения,
- журнал системного времени

может работать иначе. Мобильный протокол:

- iOS,
- Android,
- другие системы

могут иметь отдельные технические руководства. Мобильных пользователей нельзя исключать, так как захват на рабочем столе затруднён. Инструмент захвата должен адаптироваться к населению.

67. ДОСТУПНОСТЬ И ЗАХВАТ

Для участников, использующих экранные читалки, голосовой ввод или моторные вспомогательные инструменты, стандартного визуального захвата может быть недостаточно. Альтернативные доказательства:

- Экспорт дерева доступности
- Запись голосовой сессии

- Расшифровка приложения
- Журнал вспомогательных технологий
- Запись наблюдения за задачей

может происходить. Эти наблюдения не следует считать низконадёжными. Следует разработать эквивалентный стандарт доказательств.

68. НАБЛЮДЕНИЕ ЗА ИЗМЕНЕНИЕМ ПОВЕДЕНИЯ ИНСТРУМЕНТА ЗАХВАТА

Расширение браузера или инструментирование:

- загрузка страницы,
- поведение интерфейса,
- система безопасности,
- обнаружение продукта

может быть затронуто. Эффект инструмента захвата должен быть протестирован в пилотном проекте. Если сессии с инструментом и без него ведут себя по-разному: условие инструмента становится экспериментальным фактором, не можно считать инструмент захвата нейтральным молча.

69. ВЕРСИОНИРОВАНИЕ ИНСТРУМЕНТА ЗАХВАТА

Каждый пакет захвата:

- имя инструмента,
- версия,
- конфигурация,
- операционная среда,
- известные ошибки

должен содержать. Если инструмент захвата меняется во время измерения:

- новая версия инструмента,
- подволна,
- обзор совместимости

может потребоваться.

70. ОШИБКА ИНСТРУМЕНТА ЗАХВАТА

Инструмент:

- может неправильно копировать запрос,
- может обрезать конец ответа,
- может неверно зафиксировать время,
- может не сгенерировать хеш,

может сжать файл. Ошибка инструмента:

- не может быть загружен в ИИ-продукт,

- или пользователем

Отдельно:

СБОЙ_ЗАПИСИ_ИНСТРУМЕНТА

должен быть записан как.

71. СИНТЕТИЧЕСКИЙ ТРЕХВОЛНОВОЙ GEO-1000 СЛУЧАЙ

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Следующие продукты, даты, номера и результаты являются полностью вымышленными. Десять синтетических ИИ-продуктов:

A1–A10

будут измеряться в три основные волны.

71.1. Календарь волн

Волна	Тип	UTC окно поставки	Назначение
W1	Основной	12.00-13.00	Начало
W2	Подтверждающий	20.00-21.00	Краткосрочное повторение
W3	Стабильность	04.00-05.00	Среднесрочный и почасовой баланс

Для каждого продукта в каждой волне: 1,000 нацелены действительные основные наблюдения. Общая цель:

```
10×1,000×3=30,000
```

71.2. Структура слотов

Каждая волна:

- 12 микрослоты
- с участниками, сбалансированными по продукту и стране в каждом слоте

Слот участника: распределяется случайным образом до сбора данных.

71.3. Результат W1

Цель: 10,000 действительные основные наблюдения Поле:

- 13,240 приглашения
- 10,870 выполненные задания
- 10,106 действительные наблюдения с точки зрения фиксации
- Основные наблюдения анализа 10,000 в соответствии с ранее зафиксированным порядком основной/резервной записи

Оставшиеся действительные записи 106:

- резервная копия,
- чувствительность,
- контроль качества захвата

зарезервированы для.

71.4. Технические события

Лимит использования 118 Ошибка соединения 42 Отклонение системы 31 Отключение пользователя 17 Несоответствие промпта 9 Повтор скриншота 6 Все результаты первоначального контакта сохранены. Повторно обрабатывались только заранее определенные реальные технические ошибки.

71.5. Изменение продукта в период W2

Продукт A7 в 20.32 UTC:

- автоматическое использование веба,
- новый обзор ссылки

пусть побеждает. Наблюдения A7:

W2A: 20.00–20.31

W2B: 20.32–21.00

разделены как. Единственный фиксированный W2 балл для A7 не публикуется без объяснения.

71.6. Внешнее событие в W3

Важное юридическое решение должно быть опубликовано о 04.22 UTC в отношении аудированной синтетической компании. Следует отметить, что ответы в продуктах с веб-доступом изменились после 04.25. Событие: Оно добавляется в Реестр внешних событий. Результаты до и после события показаны как отдельные диагностические ячейки. Наблюдения не удаляются, поскольку они снижают или повышают балл.

71.7. Карта здоровья волны

Волна	Здоровье	Основная заметка
W1	WH-4	Сбалансированная и сильная фиксация
W2	WH-3	Подволна из-за изменения продукта A7
W3	WH-3	Запись о внешнем юридическом событии

Все три волны могут использоваться. Комментарии к A7 и внешним событиям также ограничены.

72. СЛУЧАЙ ЗАПИСИ ОДНОГО СИНТЕТИЧЕСКОГО НАБЛЮДЕНИЯ

СИНТЕТИЧЕСКИЙ ПРИМЕР

Участник:

- Турция
- Турецкий
- Платный веб-план Orion AI
- Контролируемая панель в стандартизированных условиях

W1

Слот 12.15–12.20 UTC Назначенная промпт: «С какой материнской компанией связана Apple.com, и каковы основные виды деятельности этой компании?»

72.1. Время

Открытие задания: 12.14.20 Отправка промпта: 12.16.08 Первый токен: 12.16.10 Завершение ответа: 12.16.24 Захват: 12.16.39 Генерация хэша: 12.16.41 Загрузка: 12.17.12 Принятие сервером: 12.17.13

72.2. Доказательства

Хэш запроса совпадает Полный текст ответа доступен Доступны два перекрывающихся скриншота Видны продукт, план и модель Метка памяти и специальные инструкции отключены Использование веба включено Источник не указан Копирование запрещено Предыдущие разговоры отсутствуют Доступны хэши файлов и подпись пакета Уровень захвата:

NCL-4

Срок действия:

CV-1

Пока не решено, является ли ответ правильным или неправильным.

73. СЛУЧАЙ ИСКУССТВЕННОЙ ТЕХНИЧЕСКОЙ ПОВТОРНОЙ ПОПЫТКИ

При первой попытке:

- Запрос отправлен правильно
- Продукт показал «Сетевая ошибка»
- Существенный ответ не начат
- Скриншот ошибки сделан
- Время и хеш зафиксированы

Согласно заранее определенному правилу, вторая попытка была предпринята в течение пяти минут. Ответ получен при второй попытке. Правильная запись:

- Результат первого контакта: ВРЕМЕННАЯ_ТЕХНИЧЕСКАЯ_ОШИБКА
- Первый подходящий вывод содержимого: вторая попытка
- Обе попытки включены в пакет
- Первый инцидент включен в показатель технических ошибок
- Второй ответ включается в оценку содержания, если метод предсказывает это заранее

Неверная запись: Первая ошибка удалена, и она рассматривается так, как будто существует только второй ответ.

74. СИНТЕТИЧЕСКИЙ НЕПРАВИЛЬНЫЙ СЛУЧАЙ ПОВТОРНОЙ ПОПЫТКИ

Первый ответ: «Apple — это местная компания, которая продаёт только мобильные телефоны». Участнику не понравился ответ. Он воспроизводит его снова. Второй ответ оказывается правильным. Участник отправляет только второй ответ. Журнал системы или запись захвата показывает повтор. Статус правильности:

CV-8— МАНИПУЛИРОВАННОЕ НАБЛЮДЕНИЕ

Наблюдение не сохраняется только потому, что второй ответ верен. Если первый ответ можно дополнительно восстановить, его можно сохранить в расследовании инцидента. Основная запись участника недействительна.

75. ОТЧЕТ О ЗАКРЫТИИ ВОЛНЫ

В конце каждой волны следует создать следующую запись:

- ID волны
- Время открытия и закрытия
- Целевое количество наблюдений
- Приглашенное лицо
- Выполненное задание
- Действительное наблюдение
- Наблюдение, входящее в основной анализ
- Резервное действительное наблюдение
- Техническая ошибка
- Отказ
- Лимит использования
- Несоответствие промпта
- Несоответствие захвата
- Поздняя загрузка
- Дубликат
- Подозрение на подмену
- Выход участника
- Изменения продукта
- Внешние события
- Покрытие слотов
- Охват стран и языков
- Уровни уверенности захвата
- Состояние здоровья волны
- Корень целостности
- Ответственное одобрение человеком

76. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH12-N01

Каждое наблюдение GEO-1000 должно быть связано с предварительно зафиксированной волной измерений.

CH12-N02

Термин «одновременное время» нельзя использовать без полного окна UTC, микрослот и правил завершения.

CH12-N03

Местное время не может использоваться отдельно в качестве времени волны; основным стандартом времени должен быть UTC.

CH12-N04

Окно передачи, микро-слоты, время завершения и время загрузки должны быть определены до просмотра ответов ИИ.

CH12-N05

Назначение слотов должно выполняться таким образом, чтобы не оставлять страны, языки, продукты и группы панелей несбалансированными по времени.

CH12-N06

Участников нельзя переводить на более ранние или более поздние слоты на основе их результатов.

CH12-N07

Смещение местного времени единого окна UTC должно быть задокументировано в публичной записи метода.

CH12-N08

Возврат глобальных волн или локально сбалансированные по времени операции должны сохраняться как отдельные временные результаты.

CH12-N09

Каждое указание для наблюдения должно содержать время подачи, первого контакта, завершения, захвата и загрузки в той степени, в которой это применимо.

CH12-N10

Неизвестные поля времени не могут быть представлены как предполагаемое фактическое время.

CH12-N11

Часы устройства сами по себе не могут считаться надежным доказательством времени UTC.

CH12-N12

Время захвата и время загрузки должны фиксироваться отдельно.

CH12-N13

Наблюдения, отправленные вне волны, не могут быть включены в основную волну без объяснения.

CH12-N14

Правила исключения задержки или отклонения слота должны быть независимы от содержания ответа.

CH12-N15

В результате первого контакта необходимо хранить его отдельно от первой передачи содержания.

CH12-N16

Отказ, ограничение использования, техническая ошибка, разъяснение и неопределенность не могут быть помещены в одну категорию.

CH12-N17

Неправильный, отрицательный, без ссылок или критический AI-ответ не может быть причиной повторной попытки.

CH12-N18

Техническая повторная попытка может осуществляться только при заранее определенных технических условиях.

CH12-N19

Когда выполняется техническая повторная попытка, все попытки и результат первого контакта должны быть сохранены.

CH12-N20

Неполные или частичные ответы нельзя игнорировать и заменять новым ответом.

CH12-N21

Первый подходящий результат нельзя обновлять, выбирать заново или вручную улучшать.

CH12-N22

Каждое наблюдение должно быть направлено на достижение как минимум уровня уверенности захвата NCL-3 или обоснованного эквивалента.

CH12-N23

Только скриншот не может считаться полным пакетом доказательств.

CH12-N24

Визуальный запрос доказательств должен показывать весь ответ и интерфейс физического продукта.

CH12-N25

Длинные ответы должны полностью и последовательно сохраняться с использованием метода захвата.

CH12-N26

Если существует существенная разница между исходным текстом ответа и визуальными доказательствами, должно проводиться расследование.

CH12-N27

OCR может использоваться в качестве вспомогательного метода, если прямое запись текста невозможна; без проверки это не может считаться окончательным источником текста.

CH12-N28

Начальная версия завершения и последующие материальные версии динамически изменяющихся ответов должны сохраняться отдельно.

CH12-N29

Участник, который останавливает ответ или закрывает интерфейс, должен быть зафиксирован как отдельное событие.

CH12-N30

Если запрос и ответ находятся в отдельных файлах, они должны быть связаны с одной сессией и временной цепочкой.

CH12-N31

Каждый файл доказательства и пакет доказательств должны иметь хеш целостности.

CH12-N32

Сам по себе хеш не может служить доказательством реальности наблюдения.

CH12-N33

Цифровая подпись и отметка времени не являются полной гарантией подлинности наблюдения; они являются компонентами цепочки доказательств.

CH12-N34

Никакое редактирование или корректировка не допускается прямо в исходном файле доказательства.

CH12-N35

Каждая отредактированная производная версия должна иметь собственный идентификатор файла и хеш.

CH12-N36

Доступ к исходным доказательствам может быть ограничен; публичный реестр должен демонстрировать целостность наблюдения без раскрытия личности.

CH12-N37

Имя участника, электронная почта, пароль, личная переписка или ненужная личная информация не могут быть добавлены в публичные доказательства.

CH12-N38

Необходимо также оценить риск повторной идентификации для пользователей из маленьких стран и редких подгрупп.

CH12-N39

Захват Панель пользователей в естественных условиях по умолчанию не должен собирать специальные инструкции и содержимое прошлых разговоров.

CH12-N40

Должны быть предоставлены эквивалентные альтернативные методы захвата для пользователей с особыми потребностями.

CH12-N41

Если инструмент захвата существенно изменяет опыт пользователя, этот эффект должен измеряться отдельно.

CH12-N42

Инструмент захвата и его версия должны записываться при каждом наблюдении.

CH12-N43

Если инструмент захвата существенно изменяется в течение волны, должно быть сделано разграничение для подволны или версии инструмента.

CH12-N44

Ошибка инструмента захвата не может классифицироваться как продукт ИИ или ошибка участника.

CH12-N45

Телеметрия провайдера может поддерживать независимый сбор данных пользователем; она не может его заменять без объяснения.

CH12-N46

Если во время волны происходит существенное изменение продукта AI, наблюдения не могут быть объединены как единое фиксированное состояние продукта.

CH12-N47

Изменение продукта требует разделения волны, перезапуска, явного моделирования или предупреждения о смешанном состоянии.

CH12-N48

Отсутствие объявления об обновлении от провайдера не является окончательным доказательством того, что система не изменилась.

CH12-N49

Существенные внешние события должны быть добавлены в Реестр внешних событий с указанием времени и области действия.

CH12-N50

Внешние события не могут служить причиной для исключения после получения результатов только потому, что они повысили или понизили оценку.

CH12-N51

Достоверность сбора данных должна оцениваться максимально слепо, до оценки семантической корректности ответа.

CH12-N52

Правильный ответ AI не может подтверждать отсутствие или фальсификацию сбора данных.

CH12-N53

Неправильный ответ AI не может опровергнуть сбор данных, соответствующий протоколу.

CH12-N54

Решения о дублировании и фальсификации должны включать несколько сигналов и контроль человека.

CH12-N55

Один только одинаковый текст ответа не может считаться доказательством дублирования или мошенничества.

CH12-N56

Добавление, удаление или изменение наблюдений после закрытия волны требует нового манифеста и версии анализа.

CH12-N57

Старый манифест волны и корень целостности должны быть сохранены.

CH12-N58

Сырые доказательства могут сохраняться без изменений; решения о допустимости и арбитраже могут изменяться только в версионной форме.

CH12-N59

Право участника на отзыв и этика исследований должны соблюдаться важнее утверждения о неизменности.

CH12-N60

Целостность волны должна оцениваться независимо от оценки ИИ.

CH12-N61

Отчет о закрытии волны должен показывать цель, фактические данные, аннулирования, технические события, изменения продукта и целостность доказательств.

CH12-N62

NOMOS Соответствие требованиям захвата не может быть связано исключительно с использованием программного обеспечения, разработанного NobleJackal; должны быть возможны эквивалентные открытые приложения.

CH12-N63

Схема захвата, решения валидации и записи волн должны иметь версионность в формате, читаемом машиной.

CH12-N64

Каждая волна и цепочка доказательств должны иметь ответственного человека или институционального владельца.

77. ФОРМЫ ОШИБОК

CH12-F01— ТА ЖЕ ВТОРАЯ ОШИБКА

Принуждение всех участников отправлять в одну и ту же секунду считается единственным условием одновременности.

CH12-F02— ПРИСВАИВАНИЕ "СЕГОДНЯ" КАК СИНХРОНИЗИРОВАННОЙ ВОЛНЕ

Местное время и часовые пояса путаются.

CH12-F03— ИСПОЛЬЗОВАНИЕ МЕСТНОГО ВРЕМЕНИ, КАК UTC

Наблюдения поступают в неправильном хронологическом порядке.

CH12-F04— ИЗМЕНЕНИЕ НАЗНАЧЕНИЯ СЛОТА НА ОСНОВЕ РЕЗУЛЬТАТОВ

Пользователи с низкими баллами перемещаются в другую временную ячейку.

CH12-F05— УПАКОВКА ОДНОЙ СТРАНЫ В ОДИН СЛОТ

Эффект времени смешивается с эффектом страны.

CH12-F06— ГЕНЕРАЦИЯ ИСКУССТВЕННОЙ НАГРУЗКИ

Тысячи пользователей отправляют одновременно, вызывая ненормальное поведение продукта.

CH12-F07— СКРЫТИЕ НОЧНОГО УЧАСТИЯ

Усталость и потеря участия не учитываются в некоторых областях.

CH12-F08— ДОБАВЛЕНИЕ НАБЛЮДЕНИЯ ВНЕ ВОЛНЫ К ВОЛНЕ

Ответ, полученный через часы или дни, используется так, как будто он принадлежит основному окну.

CH12-F09— УЧЕТ ВРЕМЕНИ ЗАГРУЗКИ КАК ВРЕМЕНИ СЪЕМКИ

Невозможно доказать, когда файл был сгенерирован.

CH12-F10— СЛИПЕРЕСТВЕННОЕ ДОВЕРИЕ ЧАСАМ УСТРОЙСТВА

Ручное изменение или неправильное время становится основной записью времени.

CH12-F11— ИСКЛЮЧАЯ ЗАДЕРЖАННЫЙ ОТРИЦАТЕЛЬНЫЙ ОТВЕТ

Правило времени применяется в зависимости от результата.

CH12-F12— УДАЛЕНИЕ РЕЗУЛЬТАТА ПЕРВОГО КОНТАКТА

Ограничение скорости, ошибка или отказ становятся невидимыми при второй попытке.

CH12-F13— СЧИТАЕМ ОТКАЗ ТЕХНИЧЕСКОЙ ОШИБКОЙ

Фактическое поведение системы повторяется.

CH12-F14— СЧИТАЕМ НЕПРАВИЛЬНЫЙ ОТВЕТ ТЕХНИЧЕСКОЙ ОШИБКОЙ

Восстановление выполняется до получения желаемого результата.

CH12-F15— ИГНОРИРОВАНИЕ ЧАСТИЧНОГО ОТВЕТА

Даже если система прерывается, сохраняется только последующий полный ответ.

CH12-F16— ВЫБОР ЛУЧШЕГО ОТВЕТА

Среди нескольких ответов сообщается самый положительный или самый длинный.

CH12-F17— ТОЛЬКО СКРИНШОТ

Визуального материала считается достаточно без промпта, продукта, времени и метаданных.

CH12-F18— ОТРЕЗАННЫЙ УСПЕШНЫЙ ВИЗУАЛ

Некорректные или ограничительные части остаются невидимыми.

CH12-F19— ЗАПИСЬ, НЕ ПОКАЗЫВАЮЩАЯ промпт

Невозможно доказать, к какому вопросу относится ответ.

CH12-F20— УСЕЧЕНИЕ КОНЦА ОТВЕТА

Ограничение материала или часть отказа становится невидимой.

CH12-F21— ПРОПУСК В ПОСЛЕДОВАТЕЛЬНЫХ ВИЗУАЛАХ

Части ответа теряются.

CH12-F22— ПОДСЧЕТ ТЕКСТА OCR КАК СЫРОГО ОТВЕТА

Ошибки распознавания изменяют утверждения по материалам.

CH12-F23— СЧИТАЙТЕ РЕАЛЬНОЙ ТОЛЬКО ПОСЛЕДНЮЮ ВЕРСИЮ DYNAMIC RESPONSE

Первоначальный контакт пользователя удаляется.

CH12-F24— СЧИТАЙТЕ ПОЛЬЗОВАТЕЛЬСКОЕ ПРЕРЫВАНИЕ ПОЛНЫМ ОТВЕТОМ

Системный вывод неполный из-за вмешательства пользователя.

CH12-F25— НЕПРАВИЛЬНО СОВМЕСТИТЬ ОТДЕЛЬНЫЙ ЗАПРОС И ОТВЕТ

Файлы из разных чатов связаны с одним и тем же наблюдением.

CH12-F26— ПОКЛОНЕНИЕ ХАШУ

Любой файл с хэшом считается настоящим и оригинальным.

CH12-F27— ИГНОРИРОВАТЬ ПРЕДВАРИТЕЛЬНОЕ РЕДАКТИРОВАНИЕ ХЕША

Урезанные или изменённые файлы легитимизируются с помощью хеша.

CH12-F28— СЧИТАЮ ПОДПИСЬ ДЕЙСТВИТЕЛЬНОЙ

Подписанная, но неверная запись принимается без возражений.

CH12-F29— РЕДАКТИРОВАНИЕ RAW-ФАЙЛА

Оригинальные доказательства необратимо изменены.

CH12-F30— РАССМОТРЕНИЕ ОТРЕДАКТИРОВАННОГО ПРОИЗВОДНОГО ФАЙЛА КАК ОБЫЧНОГО ДОКАЗАТЕЛЬСТВА

Публичная копия используется как оригинальный файл.

CH12-F31— СНЯТИЕ ПРИВАТНОСТИ ДЛЯ ДОКАЗАТЕЛЬСТВ

Имя участника, аккаунт и приватный чат опубликованы.

CH12-F32— РАСКРЫТИЕ УЧАСТНИКА ИЗ МАЛОЙ СТРАНЫ

Время, устройство и информация о плане позволяют идентифицировать человека.

CH12-F33— СОБИРАНИЕ СПЕЦИАЛЬНОЙ ИСТОРИИ В НАТУРАЛЬНОЙ ПАНЕЛИ

Нелишние разговоры попадают в захват.

CH12-F34— АННУЛИРОВАНИЕ ДОКАЗАТЕЛЬСТВ ДОСТУПНОСТИ

Реальный пользователь исключается, потому что стандартное визуальное представление не может быть сгенерировано.

CH12-F35— ПРЕДПОЛОЖЕНИЕ, ЧТО ИНСТРУМЕНТ ЗАХВАТА НЕПРЕДВЗЯТЫЙ

Инструмент изменяет поведение продукта, но не тестируется.

CH12-F36— СКРЫТИЕ ВЕРСИИ ЗАХВАТА

Различные версии инструмента представлены как один метод измерения.

CH12-F37— РАССМОТРЕНИЕ ОШИБКИ ИНСТРУМЕНТА ЗАХВАТА КАК ОШИБКУ ИИ

Обрезанный ответ загружается в продукт.

CH12-F38— УЧЁТ API ЛОГА НА ЭКРАНЕ ПОЛЬЗОВАТЕЛЯ

Реальный пользовательский интерфейс становится невидимым.

CH12-F39— УЧЁТ ЛОГА ПРОВАЙДЕРА КАК НЕЗАВИСИМОГО ДОКАЗАТЕЛЬСТВА

Запись измеряемой стороны становится единым доказательством.

CH12-F40— СКРЫТИЕ ОБНОВЛЕНИЯ ПРОДУКТА

Два состояния системы внутри ветки сливаются в один результат.

CH12-F41— ИГНОРИРОВАНИЕ ИЗМЕНЕНИЙ, ЕСЛИ НЕТ ОБЪЯВЛЕНИЯ

Наблюдаемое нарушение системы не рассматривается.

CH12-F42— ИСКЛЮЧЕНИЕ ВНЕШНИХ СОБЫТИЙ НА ОСНОВАНИИ РЕЗУЛЬТАТА

Ответы с негативными новостями удаляются; ответы с позитивными новостями сохраняются.

CH12-F43— СВЯЗЬ ЗДОРОВЬЯ ВОЛНЫ С ОЦЕНКОЙ

Дефектные волны с высокими оценками действительны, сильные волны с низкими оценками считаются подозрительными.

CH12-F44— ПОДСЧЕТ ПРАВИЛЬНОГО ОТВЕТА КАК ДЕЙСТВИТЕЛЬНОЙ ФИКСАЦИИ

Дефект доказательства скрывается семантическим успехом.

CH12-F45— ПОДСЧЕТ НЕПРАВИЛЬНОГО ОТВЕТА КАК НЕДЕЙСТВИТЕЛЬНОЙ ФИКСАЦИИ

Результат измерения преобразуется в ошибку данных.

CH12-F46— ПОДСЧЕТ ОДНОГО И ТОГО ЖЕ ТЕКСТА КАК ДОКАЗАТЕЛЬСТВА БОТА

Детерминированные или похожие ответы ИИ случайно становятся дубликатами.

CH12-F47— РЕШЕНИЕ О САМОСТОЯТЕЛЬНОМ ИСКАЖЕНИИ СИГНАЛА

Разница в размере файла или времени автоматически создаёт исключение.

CH12-F48— ДОБАВЛЕНИЕ НАБЛЮДЕНИЯ ПОСЛЕ ВОЛНЫ

Новые записи тихо добавляются, чтобы достичь желаемого размера выборки.

CH12-F49— УДАЛИТЬ НАБЛЮДЕНИЕ С НИЗКИМ БАЛЛОМ ПОСЛЕ ВОЛНЫ

Манифест и результат ретроспективно изменяются.

CH12-F50— УДАЛИТЬ СТАРЫЙ МАНИФЕСТ

История версий уничтожена.

CH12-F51— ИСПРАВИТЬ СЫРЫЕ ДОКАЗАТЕЛЬСТВА

Содержание файла изменено для исправления неверной классификации.

CH12-F52— ОТКАЗАТЬ В ПРАВЕ НА ОТЗЫВ

Неприкосновенность выше прав человека.

CH12-F53— ЗАКРЫТА МОНОПОЛИЯ НА ЗАПИСЬ СОБСТВЕННОЙ ПРОГРАММЫ

Допускаются только доказательства, полученные с использованием инструмента основателя.

CH12-F54— НЕ ПУБЛИКОВАТЬ СХЕМУ ЗАПИСИ

Независимая команда не может установить ту же цепочку доказательств.

CH12-F55— РАССМОТРЕНИЕ КОРНЯ ВОЛНЫ КАК ДОКАЗАТЕЛЬСТВА РЕАЛЬНОСТИ

Корень Меркла или хэш представляется так, как если бы он доказывал точность содержимого.

CH12-F56— УТВЕРЖДЕНИЕ «СНИМОК ЭКРАНА 1,000» БЕЗ ПАКЕТА ДОКАЗАТЕЛЬСТВ

Количество файлов используется как количество действительных наблюдений.

78. ПРОЦЕДУРА АУДИТА

Шаг 1— Определение цели исследования волны

Выбирается основной, подтверждающий, инцидентный, дрейфующий или повторный тип.

Шаг 2— Фиксация состояний продукта и системы ИИ

Он подключается к волновой версии реестра системы ИИ в Разделе 9.

Шаг 3— Фиксация состояний популяции, выборки и панели

Используются записи из Разделов 5–8 и Раздела 11.

Шаг 4— Фиксация реестра промптов

Версии промптов из раздела 10 привязаны к волне измерений.

Шаг 5— Установить окно волны UTC

Начало, закрытие и дополнительные продолжительности записываются.

Шаг 6— Создать микро-слоты

Участники назначаются на слоты сбалансированным образом по продукту, стране и языку.

Шаг 7— Определить источник времени

Определяются сервер, устройство и доверенные записи UTC.

Шаг 8— Зафиксировать первый контакт и правила повторной попытки

Технические и семантические состояния разделяются.

Шаг 9— Определить методы захвата

Определяются пути веб, мобильного, доступности и естественной панели.

Шаг 10— Определить цель NCL

Записать требуемый уровень уверенности захвата для основного анализа.

Шаг 11— Фиксация схемы пакета доказательств

Определены манифест, файлы, хэш и поля метаданных.

Шаг 12— Фиксация правил конфиденциальности и редактирования

Разделены необработанные, аудиторские и публичные слои.

Шаг 13— Захватите своё транспортное средство с пилотом

Изменяет ли транспортное средство поведение продукта? Верны ли записи времени и текста?

Шаг 14— Откройте реестр внешних событий

Создан процесс для мониторинга событий продукта и сущностей.

Шаг 15— Подтвердите блокировку волн

Присваивается статус WS-1 — ЗАФИКСИРОВАНО.

Шаг 16— Откройте слоты

Участникам разрешено подавать заявки только в свои слоты.

Шаг 17— Запишите результат первого контакта

Ответ, отклонение, ошибка или лимит разделены.

Шаг 18— Захват первого подходящего результата

Полный ответ сохраняется без обновления.

Шаг 19— Создание необработанных данных и метаданных

Сохраняются скриншоты, необработанный текст, состояние системы и временные метки.

Шаг 20— Выполнение локального хэша и безопасная загрузка

Файлы получают запись целостности перед загрузкой.

Шаг 21— Принятие сервером и создание хэша пакета

Цепочка доказательств закрывается временем сервера и ID пакета.

Шаг 22— Слепая проверка валидности захвата

Статус CV присваивается без знания семантического успеха.

Шаг 23— Проведение проверки на дубликаты и вмешательство

Используются множественные сигналы и ручная проверка.

Шаг 24— Исследование системы в пределах волны и внешних событий

Определите, требуется ли субволна или смешанное состояние.

Шаг 25— Назначить здоровье волны

Статус присваивается между WH-0 и WH-5.

Шаг 26— Сгенерировать манифест волны и корень целостности

Список наблюдений версионирован.

Шаг 27— Опубликовать манифест публичного доказательства

Область применения и целостность становятся видимыми без персональных данных.

Шаг 28— Утвердить отчет о закрытии волны

Цели, происшествия, дефекты, события и изменения закрываются с человеческим утверждением.

79. НЕОБХОДИМЫЕ ДОКАЗАТЕЛЬСТВА

ID волны Тип волны Версия протокола Версия регистра системы ИИ Рамка популяции Фиксация образца Фиксация состояния панели Версия реестра промптов Версия пакета ссылок Время начала и окончания UTC Список микрослот Назначение участников на слоты Источник времени Отклонения часов устройства Время передачи Время первого контакта Время завершения Время захвата Времена хэширования Время загрузки Время принятия сервером Результаты первого контакта Первые подходящие контент-выходы Технические записи повтора Частичные записи ответов Ограничения отклонений и использования Полные необработанные ответы Полные скриншоты Последовательные визуальные манифесты Записи экрана при наличии Ссылки и записи источников Настройки системы Доказательства Инструмент захвата и версия Метод захвата Уровень NCL Хэши файлов Хэш пакета

Цифровая подпись Метка времени Проверка на дубликаты Проверка на подделку Принятие решений о действительности захвата Редактирование файлов Ссылки от исходного к редактированному Манифест конфиденциальности Записи об отзыве События изменения продукта Реестр внешних событий Подразделения волны Объем слота Объем по стране и языку Технические показатели ошибок Показатель действительности захвата Состояние здоровья волны Корень целостности волны Манифест публичных доказательств Отчет о закрытии волны Журнал изменений Ответственное лицо или учреждение

80. ЧЕК-ЛИСТ АУДИТА

Тип и цель волны ясны? Были ли зафиксированы продукты ИИ и статусы системы? Определены ли версии населения, выборки и панели? Были ли зафиксированы настройки запроса до получения результатов? Окно волны UTC открыто? Были ли микро-слоты предварительно назначены? Были ли страны и языки равномерно распределены по слотам? Меняли ли участники слоты в зависимости от своих результатов? Оценивалась ли нагрузка по местному времени? Есть ли вращающиеся часы волны? Ясны ли дополнительные времена подачи и завершения? Сравнивались ли часы устройства с надежным источником UTC? Были ли разделены времена подачи, завершения, захвата и загрузки? Попали ли наблюдения вне волны в основной балл? Была ли зафиксирована результат первого контакта? Было ли техническое исключение отделено от отклонения?

Была ли повторная попытка из-за неправильного ответа? Было ли первое выполнение сохранено в технической повторной попытке? Были ли удалены частичные ответы? Является ли первый подходящий результат действительно первым результатом? Было ли использовано «Regenerate» или последующее сообщение? Показывает ли это визуальный запрос и полный ответ? В длинных ответах производится ли захват без пробелов? Соответствует ли сырой текст визуальному? Если использовался OCR, есть ли проверка человеком? Были ли сохранены динамические версии ответа? Остановил ли пользователь ответ? Привязаны ли запрос и ответ к одной и той же сессии? Зафиксированы ли инструмент и версия захвата? Изменил ли инструмент захвата поведение пользователя? Был ли указан уровень NCL? Есть ли наблюдения ниже NCL-3 в основном анализе? Есть ли у каждого файла хэш? Ясно ли, что хэш не доказывает?

Есть ли цифровая подпись или время сервера? Отдельно ли указано время съемки и время загрузки? Есть ли причина для поздней загрузки и локальный журнал целостности? Производилась ли редактировка исходных доказательств? Имеют ли отредактированные производные отдельные хэши? Раскрывает ли публичный манифест личность участника? Можно ли повторно идентифицировать пользователя из небольшой страны? Были ли частные

разговоры необоснованно зафиксированы в естественной панели? Был ли предоставлен альтернативный путь доказательств для пользователей с ограниченными возможностями? Было ли решение о дубликате основано только на одинаковом тексте? Несет ли решение о вмешательстве несколько сигналов? Знал ли валидатор захвата семантический балл? Покрыл ли правильный ответ изъян в доказательстве? Был ли неверный ответ аннулирован? Изменился ли продукт во время волны?

Была ли обработка изменения продукта выполнена как подволна или в смешанном состоянии? Были ли зафиксированы внешние события? Вызвали ли внешние события исключение согласно результату? Независит ли состояние здоровья волны от оценки? После закрытия волны были добавлены или удалены наблюдения? Сохранены ли старые манифест и корень целостности? Как обрабатывались запросы участников на выход? Существует ли публичный манифест доказательства? Есть ли явно ответственный за отчет о закрытии волны?

81. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Если все пользователи не отправят данные в одну и ту же секунду, истинная одновременность не будет достигнута.»

Подача заявки на одну секунду может создавать искусственные эффекты нагрузки, соединения и очереди. Предопределённое короткое окно UTC и случайные микро-слоты могут более надёжно измерять тот же период системы.

Возражение 2— "Продукт может измениться в течение одного часа."

Это может измениться. Поэтому требуются записи системы, отпечаток конфигурации и отслеживание изменений продукта. В случае существенного изменения волна отделяется.

Возражение 3— «Почему простого скриншота недостаточно?»

Скриншот:

- время,
- начальный вывод,
- версия промпта,
- настройка системы,
- копирование

не всегда показывать эти объекты. Требуется полный пакет доказательств.

Возражение 4— «Разве не слишком сложно проверять тысячу скриншотов по одному?»

Это тяжело. Можно использовать автоматическую проверку целостности, выборочный ручной контроль и верификацию на основе рисков. Критические или подозрительные записи проверяются более подробно.

Возражение 5— «Если существует хэш, почему нельзя сказать, что файл настоящий?»

Хэш показывает только то, что файл остался неизменным после генерации хэша. Это не показывает, как файл был создан до хэша.

Возражение 6— «Если пользователю нужно обрезать личную информацию, как будет защищён исходный файл?»

Если возможно, инструмент захвата не должен с самого начала собирать личную информацию. Если это неизбежно, исходный файл хранится в ограниченном и зашифрованном хранилище, а публичная копия сохраняется в виде отдельного редактированного производного файла.

Возражение 7— «Если ответ неправильный, почему бы нам не попробовать снова?»

Потому что неправильный ответ зависит от измерения. Повторная попытка нужна только при реальном техническом сбое.

Возражение 8— «Если система отклоняет, может ли пользователь действительно попробовать снова?»

Верно. Основная панель первого контакта сохраняет отрицательный ответ. Отдельная дорожка естественного взаимодействия может измерять последующее поведение пользователя.

Возражение 9— «Если техническая ошибка является проблемой пользовательского опыта, почему мы разрешаем повторную попытку?»

Техническая ошибка сохраняется в результате первого контакта. Могут быть выполнены заранее определенные технические повторные попытки для дополнительного исследования представительного содержания. Два результата не заменяют друг друга.

Возражение 10— «Разве собственный журнал поставщика не является самым сильным доказательством?»

Технически это может быть очень сильным доказательством. Поставщик является владельцем измеряемой системы. Оно становится более надежным при использовании вместе с независимой фиксацией пользователем.

Возражение 11— «Будет ли NOMOS Capture обязательным программным обеспечением NobleJackal?»

Нет. NOMOS Capture — это открытый протокол доказательств. Независимые инструменты, которые соответствуют эквивалентной схеме и условиям целостности, могут быть совместимыми.

Возражение 12— «Если мы не сделаем все доказательства публичными, как будет проводиться независимая проверка?»

Публичный манифест показывает целостность и масштабы. Авторизованные независимые аудиторы могут получить доступ к сокращенным или ограниченным исходным доказательствам. Конфиденциальность и проверка могут быть разработаны вместе.

Возражение 13— «Разве вся работа не будет нарушена, если важные новости появятся во время волны?»

Не всегда. Отчет является частью реальной системной среды. Его можно отделить отметкой времени до и после события. Объем оценки объясняется соответствующим образом.

Возражение 14— «Разве не проще отбросить старые ответы и продолжить работу с новой системой при обновлении продукта?»

Старые ответы актуальны на дату измерения. Их нельзя удалять. Состояния старой и новой систем должны поддерживаться как отдельные волны или подволны.

Возражение 15— «Если инструмент захвата влияет на поведение продукта, терпит ли весь метод неудачу?»

Нет. Эффект транспортного средства измеряется с пилотом. При необходимости выбирается более легкое транспортное средство, ручной захват или метод, который использует эффект транспортного средства как отдельный фактор.

Возражение 16— «Если участник отзовет свои данные, не будет ли подорвана целостность корня?»

Создана новая версия манифеста и новый корень. Старая запись может сохранять безсодержательную отметку WITHDRAWN. Она поддерживается на основе архитектуры доказательств по правам человека.

ОБЫЧНОЕ СУДЕБНОЕ РЕШЕНИЕ РАЗДЕЛА 85

В конце исследования GEO-1000 у вас может быть 30,000 скриншотов. Само по себе это число не является доказательством. Вам нужно знать: в какой волне они были созданы? К какому состоянию системы они относились? Была ли фактически отправлена та же команда? Был ли это первый результат? Был ли ответ обновлен? Изменился ли продукт в течение волны? Когда файл был зафиксирован? Когда был сгенерирован хэш? Повторялся ли идентификатор участника? Показывает ли изображение весь ответ? Совпадает ли исходный текст с изображением? Если доказательство было позже отредактировано, где оригинал? Было ли проверено захваты до семантической оценки? Вы можете кликнуть тысячу раз в течение одной секунды. Тем не менее:

- часы устройства,
- связи,
- очереди,
- состояния продукта

Это может быть по-разному. Поэтому конкуренция — это не маркетинговое шоу. Это заранее определённая временная архитектура. Когда пользователь получает неправильный ответ: сказать «Попробуйте снова» заманчиво. Но этот неправильный ответ — это реальность, которую вы пытаетесь измерить. Если вы её удалите, вы не исправляете систему. Вы портите доказательства. Когда пользователь сталкивается с реальной ошибкой подключения, техническую повторную попытку можно выполнить. Но результат первого контакта всё равно должен быть сохранён. Потому что пользователь впервые столкнулся с ошибкой. Хеш ценен. Цифровая подпись ценна. Отметка времени ценна. Скриншот ценен. Ни один из этих элементов сам по себе не содержит всей правды. Доверие:

- запрос,
- время,
- продукта,
- пользователь,
- файл,

- инициатива,
- верификация

Это возникает из совместного сохранения его цепочки. Стандарт: если сказано «Действительны только доказательства, полученные с помощью моего программного обеспечения», это ставит собственный интерес выше независимого воспроизведения. По этой причине NOMOS Capture не является брендом инструмента, а представляет собой открытую архитектуру доказательств. Независимые университеты, исследователи или аудиторские учреждения должны иметь возможность применять те же правила. Следовательно, двенадцатый закон измерений NOMOS выглядит следующим образом:

> *Конкурентность — это наблюдаемая воспроизводимость в рамках одного и того же определённого системного времени, а не в одну и ту же секунду.*

Тринадцатый закон выглядит следующим образом:

> *Первый результат не является черновиком, который можно обновлять до тех пор, пока не понравится результат.*

Четырнадцатый закон выглядит следующим образом:

> *Хэш сохраняет доказательство; он не создаёт доказательство.*

Пятнадцатый закон звучит следующим образом:

> *Действительность Capture и точность ответа независимы друг от друга.*

Его шестнадцатый закон следующий:

> *Цепочка доказательств должна делать видимой целостность наблюдения, а не личность человека.*

Его семнадцатый закон:

> *Если система изменяется во время волны, она не может действовать так, как если бы история измерений не изменилась.*

Восемнадцатый закон измерений гласит:

> *На необработанных данных ничего не написано; решения исправляются через версионирование.*

Девятнадцатый закон измерений гласит:

> *Метод захвата, который не может применяться самостоятельно, не может быть мировым стандартом.*

Порядок раздела 12 из NOMOS

> *Не говори мне «мы спросили одновременно». / Покажи, в каком окне UTC, в каком слоте и в каком состоянии системы ты спрашивал.*
>
> *Не сажайте тысячу человек в одну и ту же секунду, чтобы создать искусственную ошибку. / Разнесите время, но сохраняйте систему в одной и той же волне.*
>
> *Не смешивайте местное время друг с другом. / Не воспринимайте время устройства как абсолютную истину.*
>
> *Не сохраняйте результат первого контакта. / Сохраняйте ограничение использования, количество повторов, ошибки и уточнения отдельно.*

Не регенерируй, когда получаешь неправильный ответ. / Этот неправильный ответ — моё реальное поведение.

Ты можешь повторить попытку при технической ошибке. / Но не удаляй первую ошибку.

Не приносите только скриншот. / Принесите промпт, полный ответ, продукт, время, настройки, цепочку вмешательств и целостность файла.

Не обрезайте хорошую часть ответа. / Оставьте конец, границу и ошибку видимыми.

Не объявляйте хэш реальностью. / Докажите также то, что было до хэша.

Не редактируйте сам исходный файл. / Сохраните оригинал и создайте отдельную публичную копию.

Не публикуйте имя человека, аккаунт или приватный разговор как доказательство.

Не делайте единственного участника в маленькой стране идентифицируемым.

Не объявляйте ваш инструмент захвата нейтральным без проверки, изменил ли он меня.

Если система изменилась в середине волны, не объединяйте старые и новые ответы так, как будто они принадлежат одному продукту.

Если внешнее событие изменило систему, зафиксируйте событие. / Не удаляйте наблюдение только потому, что вам не понравился результат.

Сохраняйте доказательства неизменными. / Если решение неверное, исправьте его новой версией.

Не привязывайте стандарт захвата к закрытому программному обеспечению одной компании. / Позвольте другим установить ту же цепочку доказательств.

Сначала заблокируйте волну. / Затем распределите время. / Затем захватите первый результат. / Затем создайте его хеш. / Затем загрузите безопасно. / Затем проверьте захват без проверки его точности. / И только после этого решите, что говорит ваш ответ.

Заключительное предложение главы

Ответ в GEO-1000 становится истинной единицей измерения только тогда, когда можно показать, в каком состоянии системы, в какой момент, из какого запроса и в рамках какой неизменной цепочки доказательств он возник.

Нормативное ядро

Каждое наблюдение GEO-1000 ДОЛЖНО быть связано с заранее объявленной и версионированной измерительной волной, определяющей: - состояния ИИ-продукта, - рамки популяции и выборки, - состояния панели, - версии промптов, - окно отправки UTC, - случайные микрослоты, - сроки завершения и захвата, - правила технической повторной попытки, - схему захвата, - версии эталонных записей, - и управление владением. Синхронизация ДОЛЖНА означать наблюдение в пределах объявленного окна состояния системы UTC и присвоенной структурой слотов. Оно НЕ ДОЛЖНО представляться как идеальная миллисекундная одновременность. Каждое наблюдение ДОЛЖНО сохранять, по возможности: - время назначения, - время подачи задания, - первое системное событие, - время завершения ответа, - время фиксации, - время создания хэша, - время загрузки и - время принятия сервером. Результат первого контакта ДОЛЖЕН оставаться видимым. Техническая повторная попытка МОЖЕТ использоваться только при заранее объявленных условиях отказа инфраструктуры и ДОЛЖНА сохранять все предыдущие попытки. Некорректные, неблагоприятные, отклонённые, нецитированные, усечённые или критические результаты НЕ ДОЛЖНЫ генерироваться повторно, заменяться или исключаться просто из-за их содержания. Пакет захвата доказательств NOMOS ДОЛЖЕН сохранять точно назначенный и отправленный запрос, полный необработанный ответ, все визуальные доказательства, метаданные состояния системы, журнал попыток, временные доказательства, хэши файлов, хэш пакета, версию инструмента захвата, статус конфиденциальности и решение о допустимости захвата с ограниченным доступом. Скриншот, хэш, цифровая подпись, журнал поставщика или отметка времени НЕ ДОЛЖНЫ отдельно представляться в качестве полного доказательства подлинности наблюдения. Необработанные доказательства ДОЛЖНЫ оставаться неизменными или допускающими только добавление. Редактирования, исправления, выводы из обращения, исключения и изменения по результатам рассмотрения ДОЛЖНЫ создавать связанные версии и сохранять предыдущую запись, где это этически и законно допустимо. Материальные изменения ИИ-продукта или внешние события во время волны ДОЛЖНЫ фиксироваться по времени и обрабатываться через субволны, явное моделирование, перезапуск или предупреждения о смешанном состоянии. Достоверность захвата ДОЛЖНА определяться независимо от семантической корректности. Публичные свидетельства ДОЛЖНЫ обеспечивать проверку без ненужного раскрытия личности участников, учетных данных, истории частных разговоров или точного местоположения. Соответствие захвата NOMOS ДОЛЖНО быть реализуемым через открытые, эквивалентные системы доказательств и НЕ ДОЛЖНО зависеть исключительно от проприетарного инструмента, контролируемого основателем стандарта. Каждая волна, пакет доказательств, решение о захвате, пересмотр манифеста и запись целостности ДОЛЖНЫ версионироваться и быть атрибутированы ответственному человеку или организации.

ЧАСТЬ IV

Реальность, вынесение решения и итог

ГЛАВА 13

Пакет проверенных данных о сущности

Пакет проверенных данных о сущности

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K10; K11; K12

Граница главы

Первые двенадцать глав установили экспериментальную цепочку: один ответ не является оценкой GEO; представление — это распределение среди пользователей и продуктов ИИ; проверяемая сущность должна быть определена до того, как наблюдаются результаты; официальное представление и действительность сущности на основе доказательств должны оставаться раздельными; целевая популяция, выборка, страны и языки должны быть определены; участники должны выбираться независимо от результатов; каждый экземпляр продукта ИИ должен быть связан с реестром со всеми материальными конфигурациями; промпты должны сохранять одно и то же намерение на всех языках; контролируемые и естественные пользовательские панели должны оставаться раздельными; и каждое наблюдение должно сохраняться в синхронизированной волне и цепочке доказательств захвата NOMOS. Аудит теперь должен ответить на свой центральный вопрос:

> *По какой эталонной записи мы будем определять, является ли полученный ответ ИИ правильным, неправильным, неполным, устаревшим, неподтверждённым или выходящим за рамки?*

Раздел 4 концептуально установил два отдельных слоя реальности:

- Слой официального представления, показывающий, что само юридическое лицо говорит о себе
- Слой подтверждённой реальности юридического лица, показывающий, в какой степени существующие доказательства позволяют делать какие-либо утверждения

Этот раздел преобразует второй слой в объект операционного аудита. Имя этого объекта:

Пакет проверенных данных о сущности

Канонический английский термин:

Пакет проверенных данных о сущности

Этот операционный объект завершает третью связь в архитектуре первых двух книг: GEO Framework определяет стандарт; 99 Ошибки в GEO выявляет способы его нарушения; а система аудита измеряет и подтверждает эти нарушения. Каждый проверяемый контроль должен содержать модель или продукт, дату, страну, язык, набор запросов, количество повторений и запись измерений. В этой цепочке Truth Pack является:

«Что сказал ИИ?»

с вопросом:

«Какие доказательства позволили ему это сказать?»

— это система ссылок между ними. Этот раздел:

- что такое Truth Pack и чем он не является,
- как утверждения преобразуются в атомарные записи о реальности,
- типы источников и доказательств,
- роли ресурсов первой стороны, официальных, независимых, операционных и пользовательских,
- контрдоказательства,
- происхождение источника и иллюзия независимости,
- достаточность доказательств,
- ограничения по времени, стране и масштабу,
- утверждения открытого и закрытого мира,
- различие между «доказательств не найдено» и «доказано ложным»,
- количественные показатели, утверждения о клиентах, партнёрах, лицензиях, сертификатах и превосходстве,
- версии и предволновая фиксация Truth Pack,
- публичные и ограниченные слои доказательств,
- процессы возражений, исправлений и несогласия,
- машиночитаемый Truth Pack

определяет. В этой главе пока не рассматривается:

- все детали вынесения решений о разделении ответов ИИ на атомарные утверждения,
- согласованность между оценщиками,
- порог перехода ответа,
- критические барьеры ошибок,
- не финализирует окончательные формулы оценки NOMOS

не финализируется. Это темы следующих разделов. Ключевой вопрос главы 13:

> *Как создать Пакет Правды о субъекте, который будет оценивать ответы ИИ, зафиксированный до получения результата, содержащий как доказательства, так и контрдоказательства вместе, с четкими временными и пространственными ограничениями?*

NOMOS Вызов

Предложение на вашем сайте компании: «Мы работаем в 25 странах». Продукт ИИ сообщает пользователю то же предложение: «Компания работает в 25 странах». Правильно ли ИИ передал информацию с официального сайта? Да. Было ли утверждение фактически проверено? Мы пока не знаем. Компания присылает вам таблицу.

В таблице содержатся названия 25 стран. Однако столбцы не объяснены: в некоторых указаны только посетители веб-сайтов. В некоторых — бывшие клиенты. В некоторых — независимые дистрибьюторы. В некоторых — организованы одноразовые мероприятия. В девяти странах фактически осуществляется активное предоставление услуг. Только в трёх странах есть местное юридическое лицо.

Что означает «to operate»? Это значит иметь клиента? Создавать местную компанию? Иметь возможность предоставлять услуги удаленно? Доставлять через партнера? Или чтобы вебсайт был доступен из этой страны? Субъект, глагол и область утверждения не определены.

Следовательно, хотя есть число, достоверной реальности пока нет. Теперь рассмотрим другое утверждение: «98% наших проектов успешны». У вас есть 49 презентаций успешных проектов. Нет неудачных файлов проектов. Сколько проектов было выполнено всего? По каким критериям определяется успех? Своевременная доставка? Принятие клиентом?

Коммерческий результат? Техническая осуществимость? Удовлетворенность клиентов? Каков период? Последние три месяца? Вся история компании? Только выбранные проекты? Без знания знаменателя, 98 процентов доказать невозможно. Теперь на сайте другой компании есть такое предложение: «Мы мировые лидеры». Это предложение повторяется:

- блог компании,
- пресс-релиз компании,
- двадцать новостных сайтов, копирующих один и тот же пресс-релиз,
- пять фейковых учетных записей пользователей, открытых компанией,
- три текста, сгенерированные ИИ, суммирующие это содержание

.

Есть тридцать различных URL. Значит ли это тридцать независимых доказательств? Нет. Одна самопубликация могла отразиться на многих производных интерфейсах.

Это именно та манипуляция, которая была выявлена в предыдущем томе: скрытый текст, вымышленные пользователи и написанные самим автором комментарии, опубликованные на разных платформах, возвращаются в систему как будто это независимые доказательства. Теперь рассмотрим противоположный случай. Официальный сайт компании сообщает: «Мы предоставляем услуги только в Турции».

Однако два месяца назад компания начала подтверждённую деятельность в Германии. Контракты, счета и официальные документы подтверждают этот факт, хотя сайт не был обновлён. Искусственный интеллект отвечает: «Компания предоставляет услуги в Турции и Германии». Ответ отличается от официального сайта, но может быть ближе к материальной реальности.

Теперь рассмотрим еще более чувствительную ситуацию. На сайте медицинского учреждения написано: «Лицензированный центр». В реестре лицензия имеется у другого учреждения с похожим названием. ИИ объединяет эти два объекта и присваивает лицензию неправильному учреждению. Компания замечает эту ошибку, но не исправляет её, потому что это выгодно для них.

Здесь применяется явное положение второй книги: Ошибка, которая выгодна человеку, всё равно остаётся ошибкой. Чтобы оценить ответ ИИ, нам нужен не только список правильных предложений, но и следующая структура: О чьём утверждении идёт речь? Что именно является заявлением? В какой период времени оно действительно? В какой стране? Для какого продукта или услуги?

Действительно ли источник подтверждает утверждение? Кто контролирует источник? Другие источники говорят то же самое или они копируют один и тот же исходный источник? Есть ли контрдоказательства? Утверждение является только официальным самозаявлением? Доказательства публичные или приватные?

На момент измерения ИИ, была ли возможность доступа к этим доказательствам? Является ли результат определённым, ограниченным, спорным или неизвестным? Первое правило этого раздела:

> *Источник сам по себе не является реальностью.*

Второе положение гласит:

> *Утверждение невозможно проверить без определения его объекта, объёма, времени и объёма доказательств.*

Её третье положение гласит:

> *Разные URL не означают разные независимые доказательства.*

Четвертая статья гласит:

> *Отсутствие доказательств не является автоматическим доказательством того, что утверждение ложное.*

Пятая его статья гласит:

> *Проверенная реальность не является абсолютной и неизменной истиной; это вывод, сделанный на основании доступных доказательств в определённое время и в определённом контексте.*

Шестое положение гласит:

> *Пакет Правды не может быть создан на основе ответа ИИ. Ответ ИИ оценивается по ранее зафиксированному Пакету Правды.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — преобразовать материальные утверждения о проверяемом субъекте в справочный пакет, связанный с доказательствами и контрдоказательства, с четкими границами времени и области, не превращая их в шаблон ответа. Раздел нормализует следующие различия:

- Официальное утверждение против проверенной реальности
- Источник против доказательства
- Доказательства против поддержки утверждения
- Данные от первого лица против независимой проверки
- Независимый источник и источник, авторизованный по теме
- Разное доменное имя и разный корень доказательства
- Количество источников и независимая семья доказательств
- Прямое доказательство и производное доказательство
- Поддерживающее доказательство и встречное доказательство
- Отсутствие утверждения и ложность утверждения
- Утверждение открытого мира и запись закрытого мира
- Текущая реальность и историческая реальность
- Время, когда реальность действительна, и время публикации источника
- Публичное доказательство и ограниченное доказательство
- Доказательство, доступное ИИ, и доказательство, открытое только аудитору
- Доказанное утверждение и утверждение, доступное для публичного обнаружения
- Прямое существование и вычисленный производный результат
- Исходные данные и коэффициент производительности
- Числитель и знаменатель

- Клиент и потенциальный клиент
- Партнер и клиент
- Членство и аккредитация
- Лицензия и связь с брендом
- Сертификат и общая институциональная соответствие
- утверждение и мнение
- Мнение и сравнительное преимущество
- Жалоба и проверенное нарушение
- Расследование и конкретное преступление или ответственность
- Устаревшая запись и некорректная историческая запись
- Некорректный ответ ИИ с отсутствующим Truth Pack
- Ошибка ссылки и системная ошибка
- Целостность Truth Pack и все утверждения являются положительными
- Конфиденциальность доказательств и отсутствие доказательств
- Право субъекта на ответ и самопроверка
- Человеческое одобрение и человеческая непогрешимость
- Помощь ИИ и ИИ как окончательный авторитет истины

В конце этого раздела каждый аудит должен быть в состоянии ответить на следующие вопросы:

Какие материальные утверждения содержатся в Truth Pack?

Какому субъекту, стране, продукту и времени принадлежит каждое утверждение?

Какие доказательства это подтверждают, какие ограничивают или опровергают?

Являются ли источники действительно независимыми или они происходят из одного и того же исходного источника?

Можно ли публично проверить требование?

Доступны ли доказательства только при ограниченном доступе аудитора?

Мог ли ИИ получить доступ к этим доказательствам на момент измерения?

Является ли требование достоверным, частичным, историческим, спорным или неизвестным?

Был ли Truth Pack зафиксирован до того, как были видны ответы ИИ?

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Каждый аудит GEO-1000 должен фиксировать материальные ссылки на утверждения относительно проверяемого объекта, вместе с подтверждающими доказательствами, контрдоказательства, происхождением источника, классом доступа, объемом, юрисдикцией, сроком действия и статусом неопределенности, в версионной Пакет проверенных данных о сущности перед семантической оценкой ответов ИИ. Пакет истины:

- самостоятельно написанный вступительный файл объекта,
- презентация продаж, собирающая только положительные доказательства,
- текст мнения аудитора,

- идеальный ответ, запрошенный у ИИ,
- абсолютная реальность всего мира,
- репозиторий ссылок, состоящий только из списка URL-адресов

это не так. Пакет истины:

> *Это пакет проверочных данных, который показывает, на основании каких доказательств материальные требования в ответах ИИ могут быть подтверждены, ограничены, опровергнуты или оставлены неизвестными.*

3. ПРЕДЕЛ СЛОВА «ИСТИНА»

Слово «истина» в Truth Pack не содержит утверждения: «Этот пакет знает все абсолютные истины о сущности.» Более точное значение такова:

> *Реальная запись, основанная на доказательствах, показывающих, в указанную дату, объем и объём доказательств, в какой степени существенные решения подтверждаются.*

Пакет правды:

- подверженной изменениям при появлении новых доказательств,
- может быть оспорена,
- её охват можно сузить,
- её можно исправить, если окажется ошибочно,

Он может оставить некоторые требования неизвестными. Авторитетность Truth Pack:

- Слово «Истина» не происходит из,
- Имя NobleJackal,
- голос NOMOS,
- Автор книги

Он возникает из следующих причин:

- чёткая структура утверждений,
- надлежащие доказательства,
- контрдоказательства,
- Родословная источников,
- время,
- граница,
- имеет смысл с
- Независимый обзор,

ответственная человеческая обязанность.

4. ЧТО ТАКОЕ Пакет проверенных данных о сущности?

Пакет проверенных данных о сущности— это версионный пакет аудита, который объединяет материальные фактические записи конкретного объекта в рамках определенной даты и объема, поддерживающие и противоположные доказательства, связанные с ними, статусы источников и неразрешенные области. Представление кандидата:

```
VETP_{E,v,t}= (I, C, O, E⁺, E⁻, L, S, T, U, G)
```

Здесь:

- E: проверяемая сущность
- v: версия пакета Истины
- t: дата ссылки
- I: идентификация субъекта и граф отношений
- C: атомарные справочные утверждения
- O: Официальная запись представления
- E^+ : подтверждающие доказательства
- E^- : опровергающие доказательства и ограничивающие записи
- L: происхождение источников и граф независимости
- S: объём, страна и матрица юрисдикции
- T: записи времени и валюты
- U: неизвестные, расхождения и пробелы в справочной информации
- G: записи управления, версий, возражений и утверждений

Каждый компонент Truth Pack может иметь отдельную версию. Основная версия пакета связывает версии компонентов вместе.

5. TRUTH PACK НЕ ЯВЛЯЕТСЯ ШАБЛОНОМ ОТВЕТА

Truth Pack не является текстом, который ИИ должен произносить слово в слово. Например, Truth Pack может содержать следующее справочное утверждение: «Asteron Holdings Ltd. является владельцем бренда Asteron Travel.» ИИ может правильно сформулировать это в следующих вариантах:

- "Бренд Asteron Travel принадлежит Asteron Holdings."
- «Asteron Holdings является владельцем бренда Asteron Travel.»
- «Asteron Travel — это бренд, который принадлежит Asteron Holdings.»

Материальные отношения сохраняются. Арбитр не должен искать слово в слово совпадение. Truth Pack:

определяет правильное поле выражений

Это не накладывает единственно правильное предложение.

6. TRUTH PACK НЕ ЯВЛЯЕТСЯ БИОГРАФИЕЙ СУЩНОСТИ

О сущности может существовать бесконечное количество правильной информации. GEO-1000 Truth Pack не обязан собирать всю информацию. Пакет должен в первую очередь охватывать следующие области:

- Материальные обязанности в рамках реестра промптов
- Идентичность и оперативное ядро сущности
- Ограничения, которые изменят решение пользователя
- Лицензии, правовые вопросы, цена, услуги, страна и временные записи
- Критерии правомочности и исключения, необходимые для консультации

- Материальные утверждения, несущие риск критической ошибки
- Различия между реальностью, подтверждённой официальным представлением, и фактической реальностью

По этому принципу:

Принцип Реальности, Пропорциональный Обязанности

называется. Сбор ненужных данных:

- конфиденциальность
- стоимость,
- актуальности,
- противоречие

может увеличить риск.

7. Двенадцать ключевых компонентов пакета правды

7.1. Каноническая идентичность субъекта

Как определено в разделе 3:

- идентичность субъекта,
- тип,
- юридическое имя,
- отношения торговой марки и доменного имени,
- материнские и дочерние компании,
- временные и юрисдикционные ограничения

упакованы.

7.2. Реестр ссылочных требований

Содержит все основные требования к реальности, которые будут использоваться в аудите в атомарной структуре.

7.3. Запись официального представления

Содержит текущие и исторические заявления, опубликованные самим субъектом на канонических носителях.

7.4. Реестр доказательств

Для каждого объекта доказательств:

- источник,
- цельность,
- компетенцию,
- доступ,
- объем охвата,
- время

записываются.

7.5. Реестр контрдоказательства

Утверждение:

- опровергающее,
- ограничивающее,
- изменяющее его достоверность,
- предоставляющее другое объяснение

содержит записи.

7.6. Граф происхождения источников

Для источников:

- первый корень,
- копии,
- цитаты,
- синдикация,
- производные ИИ

показываются.

7.7. Матрица области и юрисдикции

Утверждение:

- для какой страны,
- продукта,
- услуги,
- пользователь,
- юридическое лицо

действительно.

7.8. Матрица времени и валюты

Утверждение:

- когда началось,
- когда закончилось,
- когда это было в последний раз проверено,
- какое событие могло бы это изменить

показываются.

7.9. Реестр споров и неопределенностей

Делает видимыми области, которые противоречивы, неразрешимы или имеют недостаточные доказательства.

7.10. Слой доступа к доказательствам

Доказательства:

- публичные,
- доступные для аудита,
- ограниченные,

- конфиденциальная,
- производные

как указано.

7.11. Журнал изменений и возражений

Содержит новые доказательства, исправления, встречные мнения и изменения решений в версии.

7.12. Манифест машины

Соединяет все компоненты:

- идентичность,
- имеет смысл с
- хэш,
- отношение

друг к другу.

8. АТОМНАЯ ССЫЛОЧНАЯ УТВЕРЖДЕНИЕ

Базовая единица реальности в рамках Truth Pack:

называется Атомное Ссылочное Утверждение.

Кандидатская структура:

```
C_j= (s, p, o, q, c, l, v_f, v_t, k, m)
```

Здесь:

- s: субъективная сущность
- p: предикат или отношение
- o: объект или значение
- q: квалификаторы и ограничения
- c: страна или юрисдикция
- l: язык или локаль, если материально
- vf: начало действия
- vt: окончание действия
- k: тип утверждения
- m: материализация или метка риска

Пример: Субъект: Asteron Türkiye Tourism Ltd. / Сказуемое: предоставляет услуги / Объект: управление корпоративными поездками / Страна: Турция / Тип клиента: корпоративные клиенты / Начало: январь 1, 2026/ Последняя проверка: август 18, 2026 Это предложение недостаточно атомарное: “Asteron — ведущая компания, предоставляющая многочисленные услуги по всему миру.” В этом одном предложении:

- географический охват,
- количество услуг,
- лидерство,
- идентичность компании

перепутаны вместе.

9. ТИП ЗАЯВЛЕНИЯ

Каждое атомарное утверждение должно иметь тип.

CT-01— Утверждение об идентичности

Кто является субъектом или к какому субъекту он относится.

CT-02— Утверждение о взаимоотношениях

Право собственности, контроль, партнерство, клиентство, трудоустройство или членство.

CT-03— Утверждение о деятельности и услугах

Чем занимается организация, какую услугу она предоставляет.

CT-04— Заявление о продукте

Функция продукта, версия, владелец или доступность.

CT-05— утверждение на географический охват

Страна, регион или зона обслуживания.

CT-06— Юридический и лицензионный иск

Регистрация, разрешение, лицензия, регуляторный статус.

CT-07— утверждение по цене и коммерческому сроку

Цена, налог, подписка, возврат, гарантия или оплата.

CT-08— Количественное заявление о производительности

Процент успеха, время доставки, количество пользователей, рост или эффективность.

CT-09— утверждение для клиентов и партнёров

Отношения с клиентом, спонсором, партнёром, дистрибьютором или поставщиком.

CT-10— Сертификат и утверждение на награду

Сертификат, членство, награда или аккредитация.

CT-11— Заявление о сравнительном преимуществе или превосходстве

Лучший, ведущий, быстрее, самый надежный, номер один.

CT-12— Заявление о рекомендации и пригодности

Пригодность или рекомендация для конкретного пользователя.

CT-13— Историческое заявление

Организация, прошлое владение, бывший продукт или историческое событие.

CT-14— Заявление о риске, жалобе или споре

Жалоба, расследование, суд, санкция или спор.

CT-15— Мнение или институциональная позиция

Мнение, видение или оценочное суждение, выдвинутое о самом субъекте. Тип заявления определяет, какие доказательства считаются достаточными.

10. ВАЖНОСТЬ ЗАЯВЛЕНИЯ

Не каждая правильная информация в Truth Pack оказывает одинаковое влияние на принятие решения. Существенность может быть указана как кандидат с следующими классами:

КРИТИЧЕСКИЙ-КАНДИДАТ

ГЛАВНЫЙ-КАНДИДАТ

УМЕРЕННЫЙ

КОНСУЛЬТАЦИОННЫЙ

КОНТЕКСТУАЛЬНЫЙ

Точная степень серьезности ошибки будет определена в Разделе 15. Метка в Truth Pack только:

- управляет приоритетом доказательств,
- вниманием при обзоре,
- частотой обновлений

управляет.

11. ОБЪЕКТ ДОКАЗАТЕЛЬСТВА

Каждый источник в Truth Pack должен быть записан как Объект Доказательства, а не просто как ссылка. Структура кандидата:

```
E_k= (r, a, i, p, f, s, t, x, g, h)
```

Здесь:

- r: роль источника
- a: полномочия для утверждения
- i: независимость
- p: близость к заявлению
- f: актуальность
- s: охват
- t: время
- x: класс доступа
- g: происхождение источника
- h: запись целостности

Только URL не заполняет ни одно из этих полей.

12. ОТНОШЕНИЕ ДОКАЗАТЕЛЬСТВА К ЗАЯВЛЕНИЮ

Не каждый объект доказательства поддерживает каждое заявление. Отношение между заявлением Cj и доказательством Ek может быть одним из следующих статусов:

ПОДДЕРЖИВАЕТ

ЧАСТИЧНО ПОДДЕРЖИВАЕТ

КВАЛИФИЦИРУЕТ

ПРОТИВОРЕЧИТ

УСТАРЕВАЕТ

КОНТЕКСТУАЛИЗИРУЕТ

НЕ ИМЕЕТ СУЩЕСТВЕННОГО ЗНАЧЕНИЯ

НЕИЗВЕСТНАЯ_СВЯЗЬ

Пример: Документ компании о регистрации торговой марки: может подтверждать право собственности на бренд, не подтверждает удовлетворённость клиентов. Отзыв клиента: может подтверждать опыт отдельного пользователя, не подтверждает успех 98% среди всех клиентов.

13. РОЛИ РЕСУРСОВ

SR-01— Канонический источник первой стороны

Официальные каналы, контролируемая организацией. Пример:

- Каноническая веб-страница
- Официальная политика
- Страница с ценами
- Документ о продукте

SR-02— Оперативная запись первой стороны

Записи на внутреннем или транзакционном уровне организации. Пример:

- Счет-фактура
- Контракт
- Запись о транзакции
- Запись о доставке
- Исходные данные о производительности

SR-03— Авторизованная официальная регистрация

Это реестр, который имеет юридическую или институциональную полномочность по данному вопросу. Пример:

- Реестр компаний

- Реестр лицензий
- Учреждение, выдающее сертификаты
- Судебное решение
- Регуляторная регистрация

SR-04— Двустороннее подтверждение

Это поддержка отношений обеими сторонами или по контракту. Пример:

- Взаимоотношения с клиентом
- Партнёрство
- Спонсорство
- Дистрибьюторство

SR-05— Независимый редакционный источник

Это публикация или исследование, редакционно независимое от организации. Независимость не гарантирует точность.

SR-06— Независимый аудит или исследование

Это открытая внешняя оценка его метода и объема.

SR-07— Транзакционные или Поведенческие Данные

Записи, показывающие реальные транзакции или поведение пользователей.

SR-08— Данные, Созданные Пользователем

Комментарий, жалоба, отзыв или пользовательский документ. Это индивидуальный опыт.

SR-09— Выведенные Расчёты

Ставка, классификация или сводка, полученные из необработанных записей. Формула и входные данные должны существовать.

SR-10— Синтетические Данные

Эти данные были созданы для демонстрации метода. Они не могут подтвердить утверждение о реальном объекте.

SR-11— AI-Созданный Производный Источник

Контент, созданный системой ИИ путём суммирования других источников. Он не заменяет оригинальные источники.

SR-12— Неизвестный или неразрешимый источник

Источник не был проверен или его происхождение невозможно определить. Он обладает ограниченным доверием при проверке первичного материала.

14. АВТОРИТЕТ ИСТОЧНИКА ЗАВИСИТ ОТ ТИПА УТВЕРЖДЕНИЯ

Нет универсального и единого рейтинга для всех источников. Пример: текущая публичная цена компании:

- каноническая страница с ценой,
- фактическая касса
- текущее предложение

может быть сильным источником. Для утверждения компании: «Мы мировой лидер.» на той же веб-странице показывается только самоутверждение. Это не доказывает лидерство. Для лицензии: учреждение, выдающее лицензию, более авторитетно, чем блог компании. Для опыта клиентов: один отзыв клиента — это реальное, единичное событие и не представляет всю пользовательскую аудиторию. Следовательно:

> *Качество источника не является общим; утверждение является конкретным.*

15. СЕМЬ ИЗМЕРЕНИЙ ПРОФИЛЯ ИСТОЧНИКА

Каждый источник следует оценивать по следующим критериям:

15.1. Авторитет

Авторизован ли источник давать оценку по данному типу утверждения?

15.2. Близость

Насколько источник близок к событию, процессу или записи?

15.3. Независимость

Насколько источник независим от коммерческого или институционального владельца утверждения?

15.4. Целостность

Сохранена ли аутентичность файла или записи и история их изменений?

15.5. Актуальность

Является ли источник действительным для соответствующего времени?

15.6. Масштаб

Поддерживает ли источник весь объём утверждения или только его часть?

15.7. Воспроизводимость

Может ли другой аудитор прийти к тому же суждению на основе тех же данных? Эти показатели не обязательно должны сводиться к единому «баллу источника». Источник:

- высокая авторитетность,
- низкая независимость

возможно. Другой источник:

- высокая независимость,

- отдалённость от объекта

может быть.

16. ПРАВИЛЬНАЯ РОЛЬ ИСТОЧНИКА-ПЕРВООБРАЗА

Источник-первоисточник: не является автоматически никчёмным, и не является автоматически независимым доказательством. Источники-первоисточники могут быть убедительными для утверждений, таких как:

- Цена, опубликованная организацией
- Часы поддержки
- Описание собственных продуктов и услуг
- Текст собственной политики
- Официальные заявления, раскрытые публике
- Собственные данные о транзакциях

Они могут быть недостаточны сами по себе для утверждений, таких как:

- Мировое лидерство
- Высшая удовлетворенность
- Превосходство над конкурентами
- Универсальный успех
- Независимая аккредитация
- Деловые результаты для всех клиентов

В предыдущем тексте "99 Ошибки": предупреждение «Ваш собственный сайт — это утверждение, а не доказательство» имеет центральное значение для утверждений, которые особенно требуют превосходства и независимости. Более полное нормативное заявление выглядит следующим образом:

> *Ваш собственный сайт является доказательством того, что говорит организация; он не является автоматическим доказательством независимой правильности любого утверждения относительно внешнего мира.*

17. ЧТО ТАКОЕ НЕЗАВИСИМОСТЬ?

Наличие ресурса на различных доменных именах не делает его независимым. Независимость связана со следующими аспектами: Кто создал контент? Кто принял решение о его публикации? Есть ли оплата или спонсорство? Получал ли источник текст напрямую от компании? Проводила ли публикация собственную проверку? Скопирован ли контент с другого источника? Имеет ли источник право собственности или деловые отношения с компанией? Десять материалов, опубликованных агентством на разных доменных именах, не являются десятью независимыми источниками.

18. ПРОИСХОЖДЕНИЕ ИСТОЧНИКА

Объекты доказательств должны быть связаны с происхождением источника. Кандидат на граф:

G_S = (V_S, E_S) [K10]

Здесь:

- V_S: объекты источника
- E_S: отношения цитирования, копирования, синдикации, суммирования и производные отношения

Отношения источников:

ОРИГИНАЛЬНЫЙ_ИСТОЧНИК

ЦИТАТЫ

СИНДИЦИРУЕТ

СУММИРУЕТ

ПЕРЕВОДИТ

ПРОИСХОДИТ ОТ

СГЕНЕРИРОВАНО ИИ НА ОСНОВЕ

НЕИЗВЕСТНАЯ_ПОРОДНИЦА

может быть записано как.

19. НЕЗАВИСИМАЯ СЕМЬЯ ДОКАЗАТЕЛЬСТВ

Содержимое, происходящее из одного исходного источника, может рассматриваться как одна Семья Доказательств. Пример:

- Пресс-релиз компании
- Новостной сайт, публикующий пресс-релиз дословно
- Блог, подводящий итоги новостного сайта
- Текст ИИ, подводящий итоги блога

Есть четыре URL. Возможно, что даже не существует ни одной независимой семьи доказательств. Количество источников и количество семей доказательств следует сообщать вместе.

20. ОТМЫВКА ДОКАЗАТЕЛЬСТВ

Публикация собственного заявления на других доменах, а затем использование его в качестве независимого доказательства:

Называется Отмывкой Доказательств

Пример цепочки: Компания заявляет «мы лидеры» на собственном сайте. Платный контент публикует это как «новость». Другие сайты копируют новость. Поскольку ИИ видел многие домены, он принимает утверждение как широко распространенное. Компания использует ответ ИИ как доказательство собственного утверждения. Этот замкнутый цикл:

самостоятельное заявление → производный источник → ответ ИИ → перепубликация собственного заявления

создает. Следует сделать этот цикл видимым, установив происхождение источника через Truth Pack.

21. ОГРАНИЧЕНИЕ ИСТОЧНИКА, СГЕНЕРИРОВАННОГО ИИ

Система ИИ:

- может суммировать источники,
- может сравнивать утверждения,

может генерировать новый текст. Текст ИИ:

- не является оригинальным источником,
- не может заменить исходные доказательства,

не становится независимой верификацией только потому, что другой ИИ говорит то же самое. Вывод ИИ может использоваться для следующих целей:

- направление исследований,
- извлечение утверждений,
- обнаружение конфликтов,

помощь в поиске источников. Конечный объект доказательств должен быть связан с исходным источником насколько это возможно.

22. ОГРАНИЧЕНИЯ СИНТЕТИЧЕСКИХ ДОКАЗАТЕЛЬСТВ

Синтетические данные:

- демонстрация метода,
- тестирование системы,
- верификация программного обеспечения,
- обучение арбитра

могут быть использованы для. Синтетические данные:

- фактическое количество клиентов,
- фактическая производительность,
- фактическое покрытие страны,
- фактическая лицензия

не могут быть доказаны. Даже если синтетические записи находятся в том же файле, что и реальные доказательства, они должны быть четко обозначены.

23. ДОКАЗАТЕЛЬСТВА, ПОДТВЕРЖДАЮЩИЕ, НЕДОСТАТОЧНЫ

Truth Pack должен собирать только те источники, которые подтверждают это утверждение. Для существенных требований следует активно искать следующие записи:

- Противоречивая дата
- Иной охват
- Просроченный документ
- Несостоявшаяся сделка
- Жалоба или возврат
- Предыдущие владельцы
- Лицензия, принадлежащая другому юридическому лицу
- Коррекция или отмена
- Сравнительные данные
- Недостающий знаменатель

Это обязательство называется:

Бремя контрдоказательства

Применяется всякий раз, когда существуют материальные контрдоказательства.

24. контрдоказательства

контрдоказательства не всегда полностью опровергает утверждение. Он может выполнять одну из следующих функций:

ПРОТИВОРЕЧИТ

LIMITS_SCOPE

ОГРАНИЧИВАЕТ_ВРЕМЯ

ОПРЕДЕЛЯЕТ_ИСКЛЮЧЕНИЕ

ОСПАРИВАЕТ_МЕТОД

СОМНЕВАЕТСЯ_В_АУТЕНТИЧНОСТИ

ПРЕДЛАГАЕТ_АЛЬТЕРНАТИВНЫЙ_СУБЪЕКТ

ТРЕБУЕТ_ДОПОЛНИТЕЛЬНОЙ_ПРОВЕРКИ

Пример: отрицательный отзыв на утверждение «Все клиенты удовлетворены» не обязательно полностью опровергает это утверждение. Однако он может показать, что термин «все» некорректен.

25. ХРАНЕНИЕ контрдоказательства

Запись, которая ослабляет утверждение:

- не в пользу компании,
- выглядит низкокачественно,
- нельзя тихо удалить, так как её раскрытие публике вызовет дискомфорт

Можно также зарегистрировать контрдоказательства с низкой степенью уверенности. Его статус четко указан. Арбитр придает ему соответствующий вес.

26. СТАТУСЫ ПОДДЕРЖКИ УТВЕРЖДЕНИЯ

VS-0— НЕ ОЦЕНЕНО

Утверждение ещё не было оценено.

VS-1— ТОЛЬКО ОФИЦИАЛЬНОЕ ЗАЯВЛЕНИЕ

Заявление основано только на официальной самодекларации. Оно не было опровергнуто. Дополнительной проверки не существует.

VS-2— ПОДТВЕРЖДЕНО ПЕРВОЙ СТОРОНОЙ

Заявление поддержано прослеживаемой операцией или записью транзакции первой стороны.

VS-3— ПОДТВЕРЖДЕНО ДОПОЛНИТЕЛЬНЫМИ ДОКАЗАТЕЛЬСТВАМИ

Заявление поддержано несколькими соответствующими доказательствами. Не все источники должны быть независимыми.

VS-4— НЕЗАВИСИМО ПРОВЕРЕНО

Заявление было проверено с помощью независимого или авторитетного внешнего источника, соответствующего теме.

VS-5— ПОДТВЕРЖДЕНО МНОГОИСТОЧНИКОВО

Заявление сильно поддержано различными типами и семействами источников, в объеме и с течением времени.

VS-C — ОПРОВЕРГНУТО

Более сильные или более подходящие доказательства опровергают утверждение.

VS-U — НЕРАЗРЕШЕНО

Доказательства недостаточны, противоречивы или находятся на уровне, где невозможно принять решение.

VS-W — ОТЗЫВ

Истец отозвал своё заявление. Историческая запись сохраняется.

VS-R — ОГРАНИЧЕННО ПРОВЕРЕНО

Утверждение подтверждается доказательствами, доступ к которым имеет только аудитор. Публичная проверка невозможна. Этот статус также влияет на то, можно ли ожидать, что ИИ сформирует утверждение на основе публичных источников.

27. САМО ПО СОБЕ НЕ ДОСТАТОЧНО СТАТУС ПОДДЕРЖКИ

Утверждение при этом одновременно:

- подтверждено,
- только для конкретной страны,
- только для конкретного диапазона дат,
- с доказательствами, недоступными публично,
- при противоположных мнениях

Это возможно. Следовательно, вместо одного статуса может использоваться многомерный вектор статуса реальности:

```
V_j= (поддержка, охват, время, конфликт, доступ)
```

Пример:

VS-4+ ОГРАНИЧЕННЫЙ-ОХВАТ + ТЕКУЩИЙ + БЕЗ-КОНФЛИКТА + ПУБЛИЧНЫЙ

является другим результатом.

VS-R + ГЛОБАЛЬНОЕ-утверждение + ВРЕМЕННОЙ-ОГРАНИЧЕННЫЙ + ОСПАРИВАЕМЫЙ + ОГРАНИЧЕННЫЙ

является другим результатом.

28. СТАТУСЫ ОХВАТА

SC-0— ОБЛАСТЬ НЕИЗВЕСТНА

Охват, к которому относится утверждение, определить нельзя.

SC-1— ТОЧНЫЙ ОХВАТ

утверждение и доказательства имеют одинаковый объект, продукт, страну и охват пользователей.

SC-2— ЧАСТИЧНЫЙ ОХВАТ

Доказательства подтверждают только часть утверждения.

SC-3— ЧРЕЗМЕРНАЯ утверждение

утверждение шире, чем подтверждает доказательная база.

SC-4— неверная сущность ОБЛАСТЬ

Доказательства принадлежат другому субъекту или аффилированной структуре.

SC-5— НЕВЕРНАЯ ЮРИСДИКЦИЯ

Доказательства принадлежат другой стране или юридической области.

SC-6— НЕВЕРНЫЙ ПРОДУКТ ИЛИ УСЛУГА

Доказательства принадлежат другому продукту или услуге.

29. СРОКИ ИСПОЛЬЗОВАНИЯ

TS-0— ВРЕМЯ НЕИЗВЕСТНО

Срок действия заявления неизвестен.

TS-1— ТЕКУЩИЙ

Подтверждается, что он действителен на дату ссылки.

TS-2— ИСТОРИЧЕСКИЙ

Подтверждается для конкретного прошедшего периода.

TS-3— СРОК ИСТЕК

Срок действия документа или утверждения истек.

TS-4— ЗАМЕНЕН

Он был заменен на более новую запись.

TS-5— ЗАВИСИТ ОТ СОБЫТИЯ

Он действителен до тех пор, пока продолжается событие или условие.

TS-6— БУДУЩИЙ ИЛИ ПЛАНИРУЕМЫЙ

Это план или обязательство, которое еще не произошло. Оно не может быть представлено как текущая реальность.

30. ЧЕТВЕРТАЯ ОТОМЕТКА ВРЕМЕНИ

Каждое доказательство и утверждение должны включать четыре различных времени в той степени, в какой это имеет значение:

30.1. Время, когда реальность действительна

Когда утверждение когда-либо была истинной в мире?

30.2. Время публикации источника

Когда информация была представлена публике или аудитору?

30.3. Время, когда источник был зафиксирован

Когда были зафиксированы доказательства аудита?

30.4. Время рассмотрения иска

Когда был указан статус реальности? Это различие важно. Лицензия могла быть аннулирована в 10:00. Запись могла быть обновлена в 15:00. Ответ ИИ мог быть дан в 12:00. Ответ:

- старо с точки зрения мировой реальности,
- понятно с точки зрения общедоступных доказательств

Пакет правды должен различать эти два случая.

31. ОБЩЕДОСТУПНОСТЬ ДОКАЗАТЕЛЬСТВ

Классы доступа к доказательствам:

EA-1— ОБЩЕДОСТУПНО

Общедоступно и поддаётся аудиту.

EA-2— ОБЩЕДОСТУПНО, НО ТЕХНИЧЕСКИ ОГРАНИЧЕНО

Общедоступно, но:

- доступ,
- география,
- формат,
- технический доступ

имеются ограничения.

EA-3— ДОСТУП ДЛЯ АУДИТА

Может быть предоставлен независимым аудиторам в контролируемом порядке.

EA-4— ОГРАНИЧЕННО КОНФИДЕНЦИАЛЬНО

Контракт ограничен из-за персональных данных или коммерческих тайн.

EA-5— ОБЩЕДОСТУПНОЕ РЕЗЮМЕ НА ОСНОВЕ ДАННЫХ

Сырые данные являются конфиденциальными. Их метод и резюме являются общедоступными.

EA-6— НЕДОСТУПНО ИЛИ ПОТЕРЯНО

Утверждается, что доказательства существуют, но к ним не удалось получить доступ во время аудита.

32. ОБЩЕДОСТУПНАЯ АУТЕНТИЧНОСТЬ И ОГРАНИЧЕННАЯ АУТЕНТИЧНОСТЬ

Компания может подтвердить утверждение «X является нашим клиентом» с помощью частного контракта. Контракт не может быть раскрыт общественности. Truth Pack может зафиксировать это утверждение как: VS-R — Ограниченно подтверждено. Однако нельзя ожидать, что публичный продукт ИИ будет знать об этих отношениях. Следовательно, существуют два отдельных вопроса: Правда ли утверждение? Можно ли обнаружить это утверждение публично? Утверждение может быть правдой на основе частных доказательств. Для публичного успеха GEO, опять же:

- публичные,
- его возможно потребуется проверить.
- подходящее для пользователя

может потребоваться запись.

33. ОБЩЕДОСТУПНОСТЬ

Для каждого значимого заявления может быть записано следующее поле как кандидат:

ПУБЛИЧНО_ДОСТУПНО

ПУБЛИЧНО_НО_НЕЯСНО

ПУБЛИЧНО_ПРОТИВОРЕЧИВО

ТОЛЬКО_ОГРАНИЧЕННО

НЕ_ПУБЛИЧНО

НЕИЗВЕСТНО

Реальность ответа ИИ и доступ к информации могут оцениваться отдельно впоследствии.

34. ПРАВДИВО, НО НЕ ПУБЛИЧНОЕ УТВЕРЖДЕНИЕ

Если продукт ИИ правильно сообщает информацию, которая не является публичной, возникают следующие вопросы: доступна ли информация в другом публичном источнике? Использовал ли ИИ персонализацию или частную базу знаний? Правильный ли ответ случайно? Произошла ли утечка конфиденциальной или секретной информации? Было ли утверждение проверено только задним числом? Truth Pack должен фиксировать путь источника отдельно, не считая правильность ответа сама по себе достижением.

35. УТВЕРЖДЕНИЯ ОТКРЫТОГО И ЗАКРЫТОГО МИРА

Отсутствие информации в записи зависит от типа утверждения.

35.1. Предположение открытого мира

Вся записанная вселенная неизвестна. Пример:

- Все клиенты
- Деловые связи во всех странах
- Все жалобы пользователей
- Все публикации в СМИ

Отсутствие записи не доказывает, что утверждение является ложным.

35.2. Принятие замкнутого мира

Можно предположить, что уполномоченный и исчерпывающий реестр полностью охватывает соответствующую область. Пример:

- Список действующих лицензий на конкретную дату
- Официальный статус регистрации компании
- Действительные сертификаты конкретного сертифицирующего учреждения
- Если опубликованный каталог продукции действительно дополнителен

В этом случае отсутствие записи может служить более сильным отрицательным доказательством.

36. ЗАЯВЛЕНИЕ ОБ УПОЛНОМОЧЕННОЙ ЗАМКНУТОМ МИРЕ

Если источник должен использоваться как закрытый мировой рекорд, необходимо определить следующие поля: Какую вселенную он охватывает? На какую дату? Какова частота обновления? Может ли запись отсутствовать? Существует ли архив и журнал изменений? Кто является уполномоченной организацией? Простое наличие «официального сайта» не предоставляет права закрытого мирового авторитета.

37. РАЗЛИЧИЕ МЕЖДУ НЕ НАЙДЕННЫМИ ДОКАЗАТЕЛЬСТВАМИ И ОШИБКОЙ

Следующие три предложения различны:

- «Доказательства для данного утверждения не найдены.»
- «Утверждение не подтверждается доступными доказательствами.»
- «Утверждение противоречит более сильным доказательствам.»

Первое может быть результатом исследования. Второе — это статус проверки. Третье — материальное противоречие. Это не одно и то же:

НЕ НАЙДЕНО

НЕПОДТВЕРЖДЕНО

ПРОТИВОРЕЧИТ

38. ОТСУТСТВИЕ ДОКАЗАТЕЛЬСТВ

Чтобы установить ложность утверждения на основе отсутствия данных, необходимо ответить на следующие вопросы: Является ли проверяемый универсум записей полным? Были ли проверены правильные субъекты и варианты написания имен? Были ли использованы правильные дата и страна? Могла ли запись быть обновлена с задержкой? Используется ли другое юридическое лицо или товарный знак? Есть ли проблема с доступом? Без этих условий невозможно вынести окончательное заключение, такое как «Не найдено в реестре, следовательно, лицензия не существует». Более точным результатом может быть: «Не удалось проверить в указанном реестре и на указанную дату».

39. МАТРИЦА УТВЕРЖДЕНИЙ–ДОКАЗАТЕЛЬСТВ

Тип утверждения	Основные доказательства	Роль официального сайта	Дополнительное требование
Юридическое название	Уполномоченный реестр	Указывает на заявление учреждения	Текущий реестр
Право собственности на товарные знаки	Регистрация, передача или лицензия на товарный знак	Показывает использование бренда	Проверка владельца
Текущая цена	Каноническая цена, оформление заказа, контракт	Может быть основной публичной записью	Налог и дополнительные сборы
Страну обслуживания	Контракт, операция, счет, местная запись	Публичное заявление	Определение «Услуга»
Лицензия	Лицензирующий орган	Самооценка	Объем и продолжительность
Взаимоотношения с клиентом	Контракт, счет, двустороннее утверждение	Одностороннее заявление	Дата и тип отношений
Партнёрство	Контракт или двустороннее заявление	Самооценка	Объем и продолжительность
Количество сотрудников	Персонал и кадровые записи	Приблизительное заявление	Различие между группой/компанией
Уровень успеха	Исходные данные, знаменатель, метод	Заявление о результате	Неудачные случаи
Удовлетворение	Представительный опрос и метод	Объявление результата	Отсутствие ответов
Сертификат	Учреждение, выдающее сертификат	Самооценка	Область, дата
Членство	Учреждение-член	Самооценка	Членство ≠ аккредитация
Руководство	Сравнительный метод и данные	Маркетинговое утверждение	Вселенная, критерии, дата
Рекомендация	Профиль пользователя и доказательства	Маркетинговый взгляд	Право на участие и исключения
Жалоба	Проверенная запись пользователя или учреждения	Ответ компании	Ограничение на единичный случай
Нарушение закона	Окончательная и официальная запись о статусе полномочий	Мнение учреждения	Расследование ≠ судебное решение

Эта таблица не является универсальным и неизменным рейтингом источников. Она показывает, какой тип утверждения несет какую нагрузку доказательства.

40. КОЛИЧЕСТВЕННОЕ УТВЕРЖДЕНИЕ

Количественное утверждение не может быть проверено без следующих полей:

- Измеряемый элемент
- Числитель
- Знаменатель
- Период
- Правило включения
- Правило исключения
- Источник данных

- Отсутствующие записи
- Уникальная единица
- Формула расчета
- Округление
- Дата обновления
- Ответственный человек

Пример: «98 процент успеха» отдельно не является количественным утверждением. Это неопределенное маркетинговое число.

41. ЧИСЛИТЕЛЬ И ЗНАМЕНАТЕЛЬ

Уровень успеха:

R = S/N

пусть будет. Здесь:

- S: единицы, признанные успешными
- N: все оцененные соответствующие единицы

должно быть. Если неудачные или незавершенные проекты исключаются из знаменателя, соотношение изменяется. Знаменатель нельзя выбирать после получения результата.

42. УНИКАЛЬНАЯ ЕДИНИЦА

Следующие выражения отличаются друг от друга:

- клиенты 100
- проекты 100
- счета 100
- контракты 100
- учетные записи пользователей 100
- транзакции 100

Один и тот же клиент мог выполнить десять проектов. „Проекты 100“ не означает „100 клиентов“. Truth Pack должен явно указывать единицу измерения числа.

43. НАКОПЛЕННОЕ И АКТИВНОЕ ЧИСЛО

„У нас есть 500 клиентов.“ может иметь одно из следующих значений:

- 500 уникальных клиентов за всю историю компании
- Клиенты 500 за последний год
- Активные клиенты 500 в настоящее время
- Записи контактов 500
- Счета 500
- проекты 500

Эти значения не одинаковы.

44. ВЫВОДНОЕ ЗАЯВЛЕНИЕ

Заявление может не быть прямо указано в документе. Оно может быть рассчитано из исходных данных. Пример: «Среднее время доставки составляет 12 дней». Этот результат:

- какие проекты включены,
- определение начала и конца,
- неудачные проекты,
- медианные и крайние значения

требует. Выводное заявление должно включать следующие области:

- Идентификаторы входных данных
- Формула
- Код или метод расчета
- Версия
- Запись воспроизведения

45. ЗАЯВЛЕНИЕ О ПРОИЗВОДИТЕЛЬНОСТИ

Заявление о производительности должно отвечать на следующие вопросы: Чья производительность является результатом? Какой продукт или услуга? Какой пользователь или тип проекта? Какая базовая линия? Каков критерий успеха? Каков период? Сколько неудачных результатов имеется? Является ли это выбранным случаем или всей совокупностью? Есть ли независимая проверка? Является ли результат причинным или просто корреляционным? Исследование случая: может показать возможный результат. Не показывает ожидаемый результат у всех клиентов.

46. утверждение ПО ГАРАНТИИ

Заявление «Гарантированные результаты» требует уточнения следующих аспектов: Что именно гарантируется? Доставка или бизнес-результат? Каков срок? Какие исключения существуют? Каковы условия возврата или компенсации? Есть ли третьи стороны вне контроля? Гарантируется ли консультация или рейтинг с использованием ИИ? Компания:

- может гарантировать свой собственный рабочий процесс,
- конкретная доставка

Он не может абсолютно гарантировать будущие консультационные действия ИИ-продукта, которым он не управляет.

47. утверждение КЛИЕНТА

Размещение логотипа организации на веб-сайте не является автоматическим доказательством наличия отношений с клиентом. Логотип может представлять одну из следующих форм взаимодействия:

- Клиент
- Бывший клиент
- Потенциальный клиент

- Участник мероприятия
- Спонсор
- Партнер
- Поставщик
- Пользователь технологий
- Компания группы
- Несанкционированное использование логотипа

Truth Pack должен четко определять тип отношений.

48. КОНФИДЕНЦИАЛЬНОСТЬ КЛИЕНТА

Отношения с клиентом могут быть подтверждены контрактом, но не могут быть раскрыты публично. Верная запись: VS-R — ограниченно подтверждённая. Если компания не может публично заявить: «X является нашим клиентом», ИИ не предполагается знать об этих отношениях. Количество конфиденциальных клиентов не может быть тихо преобразовано в публичный GEO-авторитет.

49. УТВЕРЖДЕНИЕ ПАРТНЕРСТВА

Слово «Партнер» само по себе недостаточно. Отношения должны включать следующие области:

- Тип партнерства
- Продукт или проект
- Страна
- Начало и конец
- Взаимное одобрение
- Коммерческий или технический охват
- Полномочия и ответственность

Программная подписка может не быть технологическим партнерством. Спонсорство мероприятия может не быть долгосрочным стратегическим партнерством.

50. УТВЕРЖДЕНИЕ ЛИЦЕНЗИИ И ПОЛНОМОЧИЙ

Утверждение лицензии должно как минимум включать следующие области:

- Полное наименование владельца лицензии
- Учреждение, выдающее лицензию
- Номер лицензии
- Страна или юрисдикция
- Объем полномочий
- Дата начала и окончания
- Статус приостановки или отмены
- Дата проверки

Лицензия другой компании группы не может быть передана бренду. Личная лицензия сотрудника не предоставляет автоматически полномочий учреждению.

51. СЕРТИФИКАТ, ЧЛЕНСТВО И АККРЕДИТАЦИЯ

Следующие понятия различаются:

- Сертификат
- Членство
- Аккредитация
- Награда
- Участие в обучении
- Знак соответствия
- Значок самооценки

Членство в профессиональной организации: может не означать экспертизу, аккредитованную этой организацией. Прохождение обучения: может не быть независимой сертификацией. Термин Truth Pack следует использовать согласно определению, данному выдающим учреждением.

52. утверждение НА НАГРАДУ

Запись о награде должна содержать следующие поля:

- Учреждение, присуждающее награду
- Год
- Категория
- География
- Заявка или структура оплаты
- Метод оценки
- Спонсорские отношения
- Побеждающая организация
- Область продукта или компании

Платная премия за заявку автоматически не является поддельной. Это также не является независимым лидерством на рынке. Статус источника должен быть сохранен.

53. ЗАЯВЛЕНИЕ О ПРЕВОСХОДСТВЕ

Выражения «Лучший.», «Мировой лидер.», «Номер один.», «Наиболее надежный.» требуют следующих областей:

- Сфера сравнения
- Набор конкурентов
- Критерии
- Вес
- Данные
- Период

- География
- Метод
- Независимость
- Неопределенность

Если эти поля отсутствуют, правильный статус Truth Pack обычно:

VS-1— Только официальное заявление

или: VS-U — Не решено.

54. РАЗГРАНИЧЕНИЕ МНЕНИЯ И ФАКТА

Это предложение: «Мы верим в дизайн, ориентированный на человека.» является корпоративным мнением или заявлением о ценностях. Для его подтверждения исследование рынка не требуется. Другое предложение: «Наши продукты лучшие в отрасли в области дизайна, ориентированного на человека.» превращается в сравнительное фактическое утверждение. Truth Pack должен сохранять эпистемический тип каждого утверждения.

55. ЖАЛОБА И ОТРИЦАТЕЛЬНАЯ ЗАПИСЬ ПОЛЬЗОВАТЕЛЯ

Жалоба пользователя:

- может быть реальным единичным событием,
- может быть неполной или неверной,
- может быть разрешена,

может принадлежать другой организации. Одиночная жалоба: «Компания притесняет всех клиентов.» не подтверждает результат. Однако серьёзное происшествие:

- Расследование,
- коррекция
- Критический риск

может быть инициировано.

56. ИСК, РАССЛЕДОВАНИЕ И ВЫНОС СУДЕБНОГО РЕШЕНИЯ

Эти статусы следует различать друг от друга:

- Иск
- Жалоба
- Открытое расследование
- Административное заключение
- Решение первой инстанции
- Окончательное решение
- Отозванное или отменённое решение
- Ответ учреждения

Наличие расследования в отношении человека или учреждения не означает, что они виновны или ответственны. Truth Pack не должен использовать язык, превышающий юридический статус.

57. ПРОПОРЦИОНАЛЬНЫЙ ЯЗЫК В СЛУЧАЯХ ОТРИЦАТЕЛЬНЫХ УТОЧНЕНИЙ

Truth Pack должен использовать следующий язык:

- "Было выдвинуто обвинение."
- "Проводится расследование указанным учреждением."
- "Указанное решение было принято в указанную дату."
- "Решение не окончательное."
- "Учреждение отрицает обвинение."
- Это не было решено в существующих записях.

Следующий язык нельзя использовать без достаточных достоверных записей:

- Является мошенником.
- Это незаконно.
- Он/Она виновен/виновна.
- Он определённо ненадёжный.

Это не просто юридический риск, это принцип точности в представлении.

58. РЕАЛЬНОСТЬ РЕКОМЕНДАЦИИ

Пакет Truth не должен сводить общую оценку компании к единственному статусу: «рекомендуется» или «не рекомендуется». Рекомендация зависит от следующих областей:

- Потребностей пользователя
- Страна
- Бюджета
- Тип услуги
- Толерантность к риску
- Вместимость
- Лицензия
- Условия исключения
- Альтернативы

Пакет правды для рекомендации:

- требования к праву на участие
- условия непригодности,
- доказанные ограничения

предоставляет. Окончательный совет пользователю является обязанностью решения продукта ИИ.

59. ЗАЯВЛЕНИЯ О ГРАНИЦАХ

Truth Pack должен содержать не только то, что он делает, но и то, чего он не делает. Пример: он работает только в Турции. Он не предоставляет юридические консультации. У него нет лицензии на здравоохранение. Цены не включают налог. Сервис доступен только корпоративным клиентам. Он не гарантирует результаты. Он не контролирует поведение некоторых третьих лиц. Информация о границах является одной из самых ценных записей для точных советов.

60. НЕИЗВЕСТНЫЙ ЖУРНАЛ

Truth Pack не должен скрывать отсутствующие поля. Типы неизвестных записей:

НЕИЗВЕСТНЫЙ_ФАКТ

НЕИЗВЕСТНЫЙ_ОБЪЕКТ

НЕИЗВЕСТНОЕ_ВРЕМЯ

НЕИЗВЕСТНАЯ_СУБЪЕКТНАЯ_ЕДИНИЦА

НЕИЗВЕСТНОЕ_ПРОИСХОЖДЕНИЕ

НЕИЗВЕСТНАЯ_ПУБЛИЧНАЯ_ДОСТУПНОСТЬ

НЕДОСТАТОЧНЫЕ_ДОКАЗАТЕЛЬСТВА

REFERENCE_CONFLICT

НЕ_ПРИМЕНИМО

ЕЩЁ_НЕ_ОЦЕНЕНО

НЕИЗВЕСТНОЕ — это не ошибка. Это защита от ложной уверенности.

61. Пробел в эталонных данных

Ответ ИИ может содержать материальное утверждение, отсутствующее в Пакете Истины при проверке. В этом случае арбитр должен использовать один из следующих вариантов:

- Утверждение выходит за рамки и не является материальным
- В Пакете Истины есть Пробел в эталонных данных
- Необходимы исследования новых доказательств
- Невозможно поддержать ответ
- Утверждение принадлежит ложной сущности

Разрыв в Пакете Истины не может быть автоматически преобразован в ошибку ИИ. Правильный статус:

РАЗРЫВ_ССЫЛКИ

может быть.

62. РАСШИРЕНИЕ ПАКЕТА ИСТИНЫ ПОСЛЕ ОТВЕТА

Ответ ИИ может создать ранее не рассмотренное материальное утверждение. Новое утверждение может быть исследовано. Однако применяются следующие правила: первая версия Пака Истины сохраняется. Новое утверждение и доказательства добавляются в отдельной версии. Начальный результат проверки не изменяется тихо. При необходимости все связанные ответы переоцениваются с новой версией. Старые и новые результаты видны в публичном журнале изменений.

63. ПРЕДВАРИТЕЛЬНАЯ ФИКСАЦИЯ ПАКА ИСТИНЫ

Перед каждой основной измерительной волной должны быть зафиксированы следующие поля:

- Версия идентичности сущности
- Версия официальной записи представления
- Регистр справок по утверждениям
- Регистр доказательств
- Реестр контрдоказательства
- Граф происхождения источников
- Матрица охвата и времени
- Регистр неизвестных
- Классы доступа к доказательствам
- Хэш пакета
- Одобрение человеком
- Время блокировки

Для этого файла:

NOMOS Пакет Истины Фиксация

может быть дано имя.

64. ПАКЕТ ИСТИНЫ И ДАТА ИЗМЕРЕНИЯ

Ответ ИИ, как правило:

состояние реальности, действительное на момент отправки запроса

должно оцениваться соответствующим образом. Пакет Истины содержит временной промежуток, который будет представлять этот момент. Реальность, которая формируется позже: не делает автоматически предыдущий ответ неверным, она создаёт новую версию Пакета Истины для новой волны.

65. ИЗМЕНЕНИЕ РЕАЛЬНОСТИ ВО ВРЕМЯ ВОЛНЫ

Компания:

- изменяет своё имя,
- получает или теряет лицензию,
- обновляет цену,
- расширяется в другую страну,
- прекращает производство продукта

тогда Пакет Истины может измениться существенно. В этом случае:

- волна измерения может быть разделена на подволны
- старые и новые версии Truth Pack могут быть связаны с отдельными ответами,

Можно использовать предупреждение MIXED_REFERENCE_STATE.

66. РЕТРОАКТИВНОЕ ИЗМЕНЕНИЕ TRUTH PACK

Если запись неверна, ее можно исправить. Однако:

- старую версию нельзя удалить,
- необходимо зафиксировать причину изменения,
- необходимо показать, какие ответы затронуты,

новый результат должен создавать отдельную версию анализа. Исправление реальности не то же самое, что переписывание истории.

67. АРХИТЕКТУРА ВЕРСИИ

Формат версии кандидата: Основная.Младшая.Патч

Основная

Изменение целевого объекта, изменение онтологии утверждения, изменение стандарта доказательств, изменение объема, которое приведет к оценке одних и тех же ответов в разных системах реальности

Второстепенная

Новое материальное утверждение Добавление новой страны или продукта Изменение статуса с новыми доказательствами Изменение исторической или текущей реальности

Патч

Орфография Метаданные Сломанная ссылка Исправление хэша или технической записи Существенное изменение реальности нельзя сохранять в виде патча.

68. АКТУАЛЬНОСТЬ И СВЕЖЕСТЬ

Не все утверждения стареют одинаково.

FRESH-1— Структурные или исторические

Поля с низкой вариабельностью, например, дата основания. Повторная валидация, связанная с событиями, может быть достаточной.

FRESH-2— зависит от события

Статус владельца, администратора или лицензии. Обязательны изменения событий и регулярные проверки.

FRESH-3— Средняя вариабельность

Страна обслуживания, количество сотрудников, партнёрство. Требуется периодическая проверка.

FRESH-4— Высокая вариабельность

Цена, наличие на складе, план, доступность продукта. Следует проверять через короткие интервалы.

FRESH-5— В реальном времени или Мгновенно

Остановка, текущее предложение, событие, статус в реальном времени. Необходимы записи, близкие к моменту измерения. Точные сроки должны определяться в зависимости от типа утверждения и отраслевого риска.

69. ДАТА ПОСЛЕДНЕЙ ПРОВЕРКИ

Каждая утверждение:

- дата последней проверки,
- рекомендуемая дата повторной проверки,
- триггер изменения

должны быть указаны. утверждение без даты последней проверки: не может считаться актуальной.

70. СНИМОК ИСТОЧНИКА

Динамическая веб-страница не должна сохраняться только как URL. Насколько это релевантно: снимок страницы,

PDF,

полный текст, скриншот, время захвата, хэш,

URL.

locale, статус доступа должен быть сохранен. Страница может измениться позже.

71. ЕСЛИ ИСТОЧНИК ПОТЕРЯН

Источник может быть удален или стать недоступным. Пакет правды:

- может сохранять снимок архива,

- хэш,
- время захвата

может быть сохранено. Но архив:

- что страница доступна публично в реальном времени,
- не может сам по себе полностью доказать объем доступа.

Ни один из этих фактов не устанавливается полностью только архивом.

72. ДОСТУП К ДОКАЗАТЕЛЬСТВАМ И СЛОИ КОНФИДЕНЦИАЛЬНОСТИ

Публичные доказательства

Доказательства, доступные общественности.

Аудиторские доказательства

Отредактированные доказательства, доступные независимым арбитрам и аудиторам.

Ограниченные доказательства

Ограниченные доказательства, содержащие контракты, персональные данные или коммерческую тайну.

Производные доказательства

Проверяемая сводка или расчет, полученные из исходных доказательств.

Публичный манифест утверждений

Для каждой утверждения к общественности:

- ее статус,
- класс источника,
- дата,
- хэш доказательства,
- статус доступа

отображение манифеста.

73. ИСПОЛЬЗОВАНИЕ КОНФИДЕНЦИАЛЬНЫХ ДОКАЗАТЕЛЬСТВ

Если необходимо использовать ограниченные доказательства: кто видел доказательство? что они проверили? сколько можно раскрыть общественности? может ли к ним получить доступ независимый второй аудитор? какова продолжительность и объем доказательств? могут ли доказательства быть отозваны? это должно быть зафиксировано. Утверждение «У нас есть конфиденциальные доказательства» само по себе не является поддержкой Truth Pack.

74. РОЛЬ АУДИРОВАННОГО СУБЪЕКТА

Аудируемое подразделение:

- может предоставлять доказательства,

- может возражать против неверных записей личности,
- может объяснять объем,
- может предложить ограниченный доступ к конфиденциальным доказательствам,

может представить контраргумент. Аудируемое лицо:

- не может присвоить окончательный подтверждающий статус своему собственному заявлению,
- нельзя удалять контрдоказательства,
- не может выбрать судью в одиночку,
- не может удалить утверждение с низкой оценкой из объема,

не может уйти от всей доказательной нагрузки, называя ее «коммерческой тайной».

75. ПРАВО НА ОТВЕТ

Прежде чем будет опубликован общедоступный материал и неблагоприятное заключение, в применимой степени аудируемое лицо имеет право на:

- соответствующее заявление,
- использованные доказательства,
- контрдоказательства,
- предлагаемый статус,
- время ответа

может быть указано. Неспособность субъекта ответить не означает, что он принимает утверждение. Ответ:

- в качестве доказательства,
- в качестве контраргумента,
- в качестве нового запроса на проверку

записываются.

76. СТАТУСЫ АПЕЛЛЯЦИЙ

NO_APPEAL (Без апелляции)

APPEAL_OPEN (Апелляция открыта)

EVIDENCE_SUBMITTED (Доказательства предоставлены)

UNDER_REVIEW (На рассмотрении)

APPEAL_UPHELD (Апелляция удовлетворена)

APPEAL_PARTIALLY_UPHELD (Апелляция удовлетворена частично)

APPEAL_REJECTED (Апелляция отклонена)

EXTERNAL_REVIEW_REQUESTED (Запрошен внешний пересмотр)

CLOSED_WITH_DISSENT (Закрыто с несогласием)

Основная запись не удаляется во время возражения. Ее статус остается видимым.

77. СТОРОННЕЕ МНЕНИЕ

Судьи или эксперты могут делать различные выводы из одних и тех же доказательств. Если существенный спор не разрешен:

- большинство решений,
- особое мнение,
- основания,
- различие в доказательствах

может быть сохранено в пакете. Пакет истины не должен создавать ложный консенсус.

78. РОЛИ УПРАВЛЕНИЯ ПАКЕТОМ ИСТИНЫ

Следующие роли могут быть определены в зависимости от актуальности:

- Владелец записи организации
- Редактор регистров утверждений
- Регистратор доказательств
- Аналитик происхождения источника
- Специалист по домену
- Юридический или нормативный рецензент
- Рецензент по количественным методам
- Рецензент контрдоказательства
- Хранитель конфиденциальности
- Арбитр пакета истины
- Рецензент апелляций

- Хранитель версии
- Ответственный человек-утверждающий

В небольшом пилотном проекте один человек может выполнять несколько ролей. Конфликты интересов должны быть раскрыты.

79. РОЛЬ ИИ В ПАКЕТЕ ИСТИНЫ

Системы ИИ могут помогать в следующих задачах:

- Извлечение утверждений из документов
- Поиск одного и того же утверждения в разных предложениях
- Оценка происхождения источника
- Отметка кандидатов на противоречие
- Удаление полей даты и области
- Предложить сопоставление доказательств и утверждений
- Обнаружить отсутствующие поля
- Создать машинную схему

AI:

- не может самостоятельно дать окончательный статус поддержки,
- не может использовать созданное им собственное резюме в качестве новых доказательств,
- не может быть единственным и безошибочным авторитетом в противоречивой реальности,

не может заменить одобрение человека.

80. УРОВНИ ПОДГОТОВКИ ПАКЕТА ИСТИНЫ

TPR-0— ПАКЕТ ОТСУТСТВУЕТ

Структурированный пакет истины не найден. Основное рассмотрение не может начаться.

TPR-1— ЧЕРНОВИК ТОЛЬКО ДЛЯ ОФИЦИАЛЬНОГО ИСПОЛЬЗОВАНИЯ

Существуют только официальные представления и самоутверждения. Может использоваться для разведки. Это не проверенный пакет истины.

TPR-2— СТРУКТУРИРОВАННЫЙ ПАКЕТ ДОКАЗАТЕЛЬСТВ

Содержит атомарные утверждения и базовые доказательства. контрдоказательства и происхождение источника могут отсутствовать.

TPR-3— ГОТОВ К АДЖУДИКАЦИИ

Утверждение, доказательства, контрдоказательства, область, время и неизвестные готовы для рассмотрения. Минимальный уровень кандидата для основного GEO-1000.

TPR-4— НЕЗАВИСИМО ПРОВЕРЕН

Пакет прошел независимую проверку на местах и по методам.

TPR-5— ВОСПРОИЗВЕДЕНО И ПОДДЕРЖИВАЕТСЯ

Пакет использовался в различных аудитах, и была проведена проверка версии и системы валюты. Основной публичный аудит как минимум:

TPR-3

должен нацеливаться на этот уровень.

81. ЭТАПЫ ПОДГОТОВКИ TRUTH PACK

Пакет должен пройти следующие этапы, прежде чем его можно будет считать TPR-3: Была ли разрешена идентичность сущности? Доступны ли необходимые поля заявлений для оперативных семейств? Являются ли материальные заявления атомарными? Ясны ли сроки и объем каждого заявления? Несут ли доказательные объекты записи целостности? Была ли проверена исходная линия? Был ли запрошен контрдоказательства? Разделён ли публичный и ограниченный доступ? Видны ли неизвестные поля? Есть ли одобрение человека и фиксация версии?

82. СООТВЕТСТВИЕ ОБЪЕМА утверждения С TRUTH PACK

Каждое семейство утверждений должно быть связано с необходимыми полями Truth Pack.

Семейство устройств	Необходимые поля Truth Pack
Основное Зеркало	Идентичность, основная деятельность, взаимоотношения сущностей
Разрешение Существования	Каноническая сущность, псевдоним, доменное имя, бренд
Активность и Объем	Продукт, услуга, страна, клиент и границы
Доказательства и Источник	Статус доказательства, происхождение источника, публичный доступ
Локальное/юрисдикционное	Местная компания, лицензия, цена, закон
Временное	Исторические и текущие записи
Рекомендация	Соответствие, исключение, риск, емкость
Сравнительное	Критерии сравнения и сопоставимые доказательства
Граница	Незавершенные задачи, страны вне рамок, ограничения гарантии

Если Truth Pack не охватывает область, требуемую заданием, эталон не готов.

83. СИНТЕТИЧЕСКИЙ APPLE.COM ОТОБРАЖЕНИЕ TRUTH PACK

СИНТЕТИЧЕСКИЙ МЕТОДОЛОГИЧЕСКИЙ ДИСПЛЕЙ / утверждения, записи и статусы ниже созданы исключительно для демонстрации метода. Они не отражают реальную информацию Apple Inc., текущую структуру компании или фактические результаты аудита. Каноническая цель: синтетическое основное корпоративное технологическое подразделение, связанное с доменом apple.com.

83.1. Синтетическая утверждение C-001

Требование:

apple.com является одной из канонических цифровые каналы, контролируемых синтетическим основным корпоративным субъектом. Тип: CT-02— Отношение / Доказательство:

- Синтетическая регистрация доменного имени
- Синтетическая официальная регистрация компании
- Синтетическая независимая проверка

Статус:

VS-4+ SC-1+ TS-1+ EA-1

83.2. Синтетическая утверждение C-002

Требование:

Основной субъект связан с потребительскими технологиями, программным обеспечением и цифровыми услугами. Тип: CT-03— Деятельность / Доказательство:

- Синтетический каталог продукции
- Синтетическая корпоративная регистрация
- Синтетическая независимая классификация

Статус:

VS-5+ SC-1+ TS-1+ EA-1

83.3. Синтетическая утверждение C-003

Требование:

Субъект является самой инновационной компанией в мире. Тип: CT-11— Превосходство / Доказательства:

- Синтетическое самоопределение компании
- Метод сравнения не определен
- Нет независимой вселенной или критерия

Статус:

VS-1+ SC-3+ TS-0+ EA-1

Правильная форма ИИ: «Компания позиционирует себя как инновационная технологическая компания.» Неправильное эпистемическое утверждение: «Это самая инновационная компания в мире.»

83.4. Синтетическая утверждение C-004

Требование:

Цены одинаковы во всех странах.

Доказательства:

Разные ценовые и налоговые условия на страницах синтетических стран

Статус:

VS-C + SC-5+ TS-1

83.5. Синтетическая утверждение C-005

Требование:

Субъект представляет конкретные продукты на указанных рынках.

Статус:

VS-4+ SC-2+ TS-1

Утверждение без указания страны: «Предлагается во всем мире.» не может быть расширено таким образом.

84. КОНТЕЙНЕР СИНТЕТИЧЕСКИХ ПАКЕТОВ ASTERON TRUTH

СИНТЕТИЧЕСКИЙ СЛУЧАЙ — НЕ РЕАЛЬНОЕ УЧРЕЖДЕНИЕ

Структура субъекта:

- Asteron Holdings Ltd. — материнская компания
- Asteron Travel — бренд
- Asteron Turkey Tourism Ltd. — местная торговая компания
- Mira Destination Services — независимый партнер по доставке
- Asteron Club Antalya — франшиза

84.1. Утверждение A-001

«Asteron работает в 25 странах.» Доказательства:

- Официальный сайт: 25 стран
- Операционная запись: активная доставка в странах 9
- Запись о бывших клиентах: исторические клиенты в дополнительных странах 16
- Местная компания: страны 3
- Принятие удаленного предложения: страны 14

Проблема: определение «Активность» неясно. Truth Pack разбивает утверждение:

A-001 a

У Asteron есть как минимум один исторический клиент или коммерческий контакт в странах 25. Статус:

VS-3+ TS-2+ SC-1

A-001 b

Asteron предоставляет активные услуги в 9 странах по состоянию на отчетную дату. Статус:

VS-4+ TS-1+ SC-1

A-001 c

Asteron имеет местное юридическое представительство в 25 странах. Статус:

VS-C + SC-3

Без этой детализации ответ ИИ «25 стран» не может быть справедливо оценен.

84.2. Утверждение A-002

"98% из проектов успешны." Представленные доказательства:

- 49 положительных случаев
- 1 неудачных случаев
- Нет полного списка проектов
- Нет определения успеха

Статус Truth Pack: VS-U — Недостаточный знаменатель и метод. Правильный формат ИИ: «Компания сообщает о высоком успехе в выбранных случаях; метод и полный знаменатель не могут быть проверены для общей ставки 98%».

84.3. Утверждение A-003

«Asteron — это организация мирового стандарта с независимой проверкой». Доказательство: стандарт был опубликован самой Asteron. Независимой проверки не найдено. Сотрудник университета поделился текстом со своего личного аккаунта. Университет не давал институционального одобрения. Статус: VS-C — Заявление о независимой проверке противоречит фактам. Правильное утверждение: «Asteron — это организация, которая публикует собственный стандарт.»

84.4. Утверждение A-004

«Все проекты гарантированы.» Контракт: Гарантирована поставка исходных файлов. Коммерческие результаты не гарантированы. Консультации ИИ не гарантированы. Атомы Truth Pack: Конкретные поставки гарантированы по контракту. Коммерческие результаты не гарантированы. Поведение ИИ третьих сторон не гарантировано. Если ИИ говорит: «Все проекты гарантированно дают результаты.» происходит расширение материального объема.

84.5. Утверждение A-005

«Asteron Club Antalya управляется напрямую Asteron Holdings.» Доказательства:

- Франчайзинговое соглашение
- Отдельный местный оператор
- Лицензия на бренд

Статус:

VS-C + SC-4

Правильное соотношение: Asteron Club Antalya управляется независимым франчайзинговым оператором по лицензии на бренд.

85. ПУБЛИЧНАЯ КАРТА РЕЗУЛЬТАТОВ TRUTH PACK

Публичная карта кандидата должна включать следующие поля:

- ID Truth Pack
- Идентичность сущности
- Дата ссылки
- Версия
- Область утверждения
- Номер атомарного утверждения
- Количество общедоступных утверждений с доказательствами
- Число утверждений с ограниченной доказательной базой
- Количество утверждений, основанных исключительно на официальной само-декларации
- Количество противоречивых утверждений
- Количество неизвестных утверждений
- Количество устаревших утверждений
- Заявления с контрдоказательства
- Открытые возражения
- Последнее обновление
- Дата повторной проверки
- Хэш целостности пакета
- Ответственное лицо или организация

Эта карточка: не указывает, что все утверждения правдивы; она показывает видимость статуса правдивости утверждений внутри пакета.

86. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH13-N01

Каждое основное решение GEO-1000 должно основываться на версии Пакет проверенных данных о сущности.

CH13-N02

Ответы ИИ Truth Pack должны быть зафиксированы перед семантической проверкой.

CH13-N03

Truth Pack не может быть представлен как абсолютное или неизменное утверждение о реальности.

CH13-N04

Truth Pack должен быть связан с конкретным субъектом, датой, страной, продуктом, услугой и областью запроса.

CH13-N05

Официальное заявление сущности и реальности, подтвержденной доказательствами, должно храниться как отдельные записи.

CH13-N06

Первичный источник не может считаться бесполезным только потому, что он первичный.

CH13-N07

Первичный источник автоматически не считается доказательством внешнего превосходства или независимости.

CH13-N08

Количество различных URL-адресов нельзя использовать как число независимых доказательств.

CH13-N09

Происхождение, синдикация, цитирование, перевод и вывод источников с помощью ИИ должны быть записаны в графе родословной источников в той мере, в какой это имеет значение.

CH13-N10

Источники, полученные из одного и того же исходного утверждения, не могут считаться независимыми доказательствами без пояснения.

CH13-N11

Сводка, созданная ИИ, не может использоваться в качестве независимого доказательства вместо исходных источников.

CH13-N12

Синтетические данные не могут использоваться для подтверждения утверждения о реальном объекте.

CH13-N13

Каждое утверждение в пакете Truth Pack должно содержать поля атомарного субъекта, предиката, значения, области применения, страны и времени, в той мере, в какой это релевантно.

CH13-N14

Составные и неоднозначные маркетинговые предложения не могут использоваться в качестве атомарных записей о реальности.

CH13-N15

Каждое доказательство должно явно показывать, какое утверждение оно поддерживает, ограничивает или опровергает.

CH13-N16

Авторитетность, близость, независимость, добросовестность, своевременность и охват источника по данной теме должны оцениваться в зависимости от типа утверждения.

CH13-N17

Одна лишь независимость не может считаться гарантией точности или авторитета.

CH13-N18

Официальный статус сам по себе не может считаться независимой проверкой утверждения о внешнем мире.

CH13-N19

Пакет истины не может быть сформирован исключительно из поддерживающих доказательств; материальный контрдоказательства должен быть активно запрошен и сохранён.

CH13-N20

контрдоказательства не может быть исключён просто потому, что он не в пользу проверяемой сущности.

CH13-N21

Отсутствие доказательств не может использоваться как эквивалент ложности утверждения.

CH13-N22

СТАТУСЫ НЕ НАЙДЕНО, НЕПОДТВЕРЖДЕНО, НЕРАЗРЕШЕНО и ПРОТИВОРЕЧИВО должны сохраняться как отдельные статусы.

CH13-N23

Доказательства отсутствия следует использовать только тогда, когда ясно определены полномочия, охват, сроки и взаимодополняемость соответствующего универсального набора записей.

CH13-N24

В утверждениях открытого мира отсутствие записей не может автоматически рассматриваться как доказательство ложности.

CH13-N25

Источники, утверждающие наличие закрытого мирового набора записей, должны объяснить охват и условия обновления источника.

CH13-N26

Время действительности утверждения, время публикации источника, время фиксации и время вынесения решения должны фиксироваться отдельно.

CH13-N27

Текущие и исторические утверждения не могут объединяться без объяснения.

CH13-N28

Просроченные, измененные или отозванные доказательства не могут использоваться как текущая реальность.

CH13-N29

Каждое утверждение должно содержать дату последней проверки и, в соответствующих случаях, триггер для повторной проверки.

CH13-N30

Публично доступные доказательства и ограниченные доказательства аудитора должны иметь отдельные статусы доступа.

CH13-N31

Утверждение, проверенное с использованием ограниченных доказательств, не может считаться информацией, доступной для публичного ИИ-продукта.

CH13-N32

Заявление «существуют конфиденциальные доказательства» не считается подтвержденным без доступа аудитора и записи целостности.

CH13-N33

Истинность утверждения и его доступность для общественности должны оцениваться отдельно.

CH13-N34

Количественное утверждение не может быть проверено без указания числителя, знаменателя, периода, единицы измерения, включений, исключений и метода расчета.

CH13-N35

Выбранные положительные примеры нельзя представлять как общий показатель эффективности.

CH13-N36

Номера клиентов, проектов, контрактов, транзакций, пользователей и счетов нельзя использовать взаимозаменяемо.

CH13-N37

Накопительные и активные показатели клиентов или пользователей следует держать раздельно.

CH13-N38

Исходные данные для заявлений должны содержать формулу и версию расчета.

CH13-N39

Отношения между клиентами и партнерами нельзя проверить без информации о типе отношений, их охвате и времени.

CH13-N40

Использование логотипа само по себе не доказывает наличие отношений с клиентом или партнером.

CH13-N41

Особенности группы, дочернего предприятия, франшизы, дистрибьютора и бренда нельзя передавать друг другу без явных доказательств.

CH13-N42

Лицензия и сертификат должны быть связаны с правильным юридическим лицом, страной, областью и информацией о дате.

CH13-N43

Членство, сертификация, аккредитация, участие в обучении и награда не могут использоваться взаимозаменяемо.

CH13-N44

Заявления о превосходстве и лидерстве не могут считаться проверенными фактами без указания сравнения, критериев, метода, даты и объема.

CH13-N45

Мнение компании и маркетинговая позиция не могут быть подняты до уровня независимого факта.

CH13-N46

Жалоба, расследование, административное заключение и окончательное решение должны различаться друг от друга.

CH13-N47

Незавершенные негативные записи не могут представляться в терминах определенного преступления, мошенничества или юридической ответственности.

CH13-N48

Контрзаявление и апелляция проверяемого субъекта должны быть связаны с соответствующей записью утверждения.

CH13-N49

Отсутствие ответа со стороны субъекта не означает его согласие с утверждением.

CH13-N50

Truth Pack должен явно сохранять неизвестные и ссылки на пробелы.

CH13-N51

Новая утверждение ИИ, отсутствующая в Truth Pack, не может автоматически считаться ложной; необходимо провести проверку REFERENCE_GAP.

CH13-N52

Запись о реальности, добавленная после ответа ИИ, должна создать новую версию Truth Pack.

CH13-N53

Новая версия Truth Pack не может тихо изменять старую историю вынесения решений и оценок.

CH13-N54

Если материальная реальность меняется в ходе волны, соответствующие наблюдения должны быть связаны с правильной временной рамкой Truth Pack.

CH13-N55

Изменения в Truth Pack должны регистрироваться в соответствии с логикой основных, второстепенных или патч-версий.

CH13-N56

Изменения в материальной реальности или статусе доказательств не могут скрываться под патч-исправлениями.

CH13-N57

Исходные доказательства и старые версии Truth Pack должны сохраняться в пределах этических и правовых норм.

CH13-N58

Truth Pack не может полагаться исключительно на документы, предоставленные проверяемым объектом.

CH13-N59

Проверяемый объект не может определять окончательный статус поддержки своих собственных утверждений.

CH13-N60

Роли подготовки и утверждения окончательного статуса в Truth Pack должны быть разделены насколько это возможно.

CH13-N61

Инструменты ИИ могут помогать в выводах и обнаружении противоречий; они не могут самостоятельно определять окончательный статус реальности.

CH13-N62

Когда Truth Pack становится общедоступным, личная информация, коммерческие тайны и ограниченные контракты не могут быть опубликованы без необходимости.

CH13-N63

Если используются ограниченные доказательства, должен быть доступен независимый или соответствующий путь второго рассмотрения.

CH13-N64

Truth Pack должен показывать общественности статусы заявлений манифеста, классы доступа, версию и записи целостности.

CH13-N65

Для первичного рассмотрения Truth Pack должен иметь как минимум TPR-3— Готовность к рассмотрению или обоснованный эквивалентный уровень.

CH13-N66

Уровень подготовки Truth Pack должен оцениваться независимо от оценки ИИ.

CH13-N67

Отсутствие Truth Pack не может использоваться как инструмент для прохождения или провала ответа ИИ в соответствии с желаемым результатом.

CH13-N68

Каждая версия пакета Truth Pack должна иметь человека или организацию-владельца, несущего ответственность.

87. ФОРМЫ ОШИБОК

CH13-F01— ОТНОСИТЕЛЬНО САЙТА КОМПАНИИ КАК К АБСОЛЮТНОЙ РЕАЛЬНОСТИ

Все самозаявленные данные рассматриваются как проверенные факты.

CH13-F02— ПОЛНОСТЬЮ ОТКАЗЫВАЯСЬ ОТ ДОКАЗАТЕЛЬСТВ ПЕРВОЙ СТОРОНЫ

Прямые институциональные записи, такие как цены, политика и транзакции, игнорируются, поскольку считаются самосгенерированными.

CH13-F03— ПОДСЧЁТ КОЛИЧЕСТВА URL В КАЧЕСТВЕ КОЛИЧЕСТВА ДОКАЗАТЕЛЬСТВ

Двадцать копий одного и того же пресс-релиза считаются двадцатью независимыми доказательствами.

CH13-F04— ОТБЕЛИВАНИЕ ДОКАЗАТЕЛЬСТВ

Самодекларация повторяется на других сайтах и используется как независимое подтверждение.

CH13-F05— ПОДСЧЁТ ЭХО ИИ КАК ДОКАЗАТЕЛЬСТВА

Утверждение считается проверенным, потому что несколько систем ИИ повторяют одно и то же утверждение.

CH13-F06— ДОБАВЛЕНИЕ СИНТЕТИЧЕСКИХ ДАННЫХ К РЕАЛЬНЫМ ДОКАЗАТЕЛЬСТВАМ

Образец метода входит в знаменатель реальной производительности.

CH13-F07— ПОДСЧЁТ КОМПОУНДНОГО МАРКЕТИНГОВОГО ЗАЯВЛЕНИЯ КАК АТОМНОГО УТВЕРЖДЕНИЯ

Идентичность, лидерство, деятельность и страна объединяются в одном утверждении.

CH13-F08— ИСПОЛЬЗОВАНИЕ ДОКАЗАТЕЛЬСТВ неверная сущность

Доказательства от материнской компании, дочерней компании или сотрудника передаются целевой организации.

CH13-F09— НЕПРАВИЛЬНАЯ СТРАНА ИЛИ ОБЛАСТЬ ПРОДУКТА

Лицензия в одной стране обобщается на весь мир, сертификат продукта обобщается на весь ассортимент.

CH13-F10— УЧЁТ УСТАРЕВШЕЙ ЗАПИСИ КАК АКТИВНОЙ

Используются старая цена, старый сотрудник или истекшая лицензия как текущие.

CH13-F11— ПРИНИМАНИЕ ДАТЫ ПУБЛИКАЦИИ ИСТОЧНИКА ЗА ДАТУ СОБЫТИЯ

Публикация документа путается с временем наступления события.

CH13-F12— УЧИТЫВАЯ ПОЗЖЕ ОПУБЛИКОВАННЫЕ ДОКАЗАТЕЛЬСТВА, КАК БУДТО ОНИ БЫЛИ ДОСТУПНЫ РАНЕЕ

Информация, недоступная на дату измерения ИИ, оценивается так, как если бы она была открыта в системе.

CH13-F13— УЧИТЫВАЯ ОТСУТСТВИЕ ДОКАЗАТЕЛЬСТВ КАК НЕПРАВИЛЬНОЕ

Заявление, не найденное в исследовании, считается опровергнутым.

CH13-F14— РАССМАТРИВАЯ ОТКРЫТЫЙ МИР КАК ЗАКРЫТЫЙ МИР

Неполный список клиентов или партнёров используется как дополнительный реестр.

CH13-F15— НЕ ПРОВЕРЯЯ ЗАПИСИ ЗАКРЫТОГО МИРА

Доказательство отсутствия предоставляется без знания охвата и актуальности реестра.

CH13-F16— НЕ ПОИСК контрдоказательства

Используются только положительные документы, предоставленные заявителем.

CH13-F17— ТИХОЕ УДАЛЕНИЕ контрдоказательства КАК НИЗКОГО КАЧЕСТВА

Записи с низкой достоверностью уничтожаются без присвоения статуса.

CH13-F18— ЧРЕЗМЕРНОЕ ОБОБЩЕНИЕ контрдоказательства

Один негативный случай считается достаточным для того, чтобы дискредитировать все достижения.

CH13-F19— ПРОЦЕНТ БЕЗ ЗНАМЕНАТЕЛЯ

Уровень успеха или удовлетворенности используется так, как если бы это была универсальная истина.

CH13-F20— ПРИНИМАЕТ ВЫБРАННЫЙ СЛУЧАЙ ЗА ОБЩЕЕ ДОСТИЖЕНИЕ

Только успешные проекты генерируют тарифы.

CH13-F21— УЧЕТ ПРОЕКТА КАК КЛИЕНТА

Несколько проектов от одного и того же клиента учитываются как отдельные клиенты.

CH13-F22— ПРЕВРАЩЕНИЕ НАКОПЛЕННОГО ЧИСЛА В АКТИВНОЕ ЧИСЛО

Количество клиентов, приобретенных за всю историю компании, представлено как текущие активные клиенты.

CH13-F23— СКРЫТИЕ ФОРМУЛЫ ВЫВЕДЕННОГО ЧИСЛА

Среднее значение или ставка не может быть воспроизведена.

CH13-F24— ИЗМЕНЕНИЕ ЗНАМЕНАТЕЛЯ ПОСЛЕ РЕЗУЛЬТАТА

Подходящие проекты выбираются в соответствии с желаемой успешностью.

CH13-F25— УЧЁТ ЛОГОТИПА КАК ДОКАЗАТЕЛЬСТВА КЛИЕНТА

Деятельность, использование технологии или несанкционированного логотипа представляется как доказательство со стороны клиента.

CH13-F26— УЧЁТ СЛОВА «ПАРТНЕР» КАК АВТОРИТЕТА

Простое владение отношениями превращается в полномочия по аккредитации или представлению.

CH13-F27— УЧЁТ ЧЛЕНСТВА КАК АККРЕДИТАЦИИ

Членство в организации используется как независимое подтверждение качества.

CH13-F28— УЧЁТ УЧАСТИЯ В ОБУЧЕНИИ КАК СЕРТИФИКАЦИИ

Участие в семинаре превращается в профессиональную лицензию.

CH13-F29— УЧЁТ ПЛАТНОЙ НАГРАДЫ КАК НЕЗАВИСИМОГО ЛИДЕРСТВА

Структура и метод вознаграждения скрыты.

CH13-F30— ПРИНИМАНИЕ СОБСТВЕННОГО СТАНДАРТА КАК НЕЗАВИСИМОГО СТАНДАРТА

Представлено как внешняя аккредитация, подобная методу, опубликованному компанией.

CH13-F31— РАССМОТРЕНИЕ ОГРАНИЧЕННЫХ ДОКАЗАТЕЛЬСТВ КАК ОБЩЕДОСТУПНЫХ

Частный контракт оценивается как открытый источник, доступный ИИ.

CH13-F32— ЗАЯВЛЕНИЕ «СУЩЕСТВУЮТ СЕКРЕТНЫЕ ДОКАЗАТЕЛЬСТВА»

Утверждение принимается без того, чтобы доказательства были увидены и проверены.

CH13-F33— РАССМОТРЕНИЕ ПРАВДЫ, ЗАКРЫТОЙ ДЛЯ ОБЩЕСТВЕННОСТИ, КАК ГЕО-УСПЕХ

Особая информация, которую ИИ не может знать, представлена как публичная сила бренда.

CH13-F34— РАССМОТРЕНИЕ МНЕНИЯ КАК ФАКТА

Фраза «Мы лучшие» становится независимой реальностью.

CH13-F35— РАССМОТРЕНИЕ ЖАЛОБЫ КАК ОПРЕДЕЛЕННОГО НАРУШЕНИЯ

Одно заявление пользователя представляется как юридическое последствие.

CH13-F36— РАССМОТРЕНИЕ РАССЛЕДОВАНИЯ КАК ПРЕСТУПЛЕНИЯ

Не завершенный процесс превращается в окончательный приговор.

CH13-F37— ИСКЛЮЧЕНИЕ НЕГАТИВНОЙ ЗАПИСИ ИЗ СФЕРЫ

Противоположная запись удаляется для защиты публичной репутации.

CH13-F38— СОХРАНЕНИЕ БЛАГОПРИЯТНОЙ ОШИБКИ

Поскольку это AI-институт, когда он отображается как большой, ограничение на Truth Pack не применяется.

CH13-F39— СЧИТАЕТ ИНФОРМАЦИЮ О ГРАНИЦАХ НЕСУЩЕСТВЕННОЙ

Невыполненные задачи и недопустимые пользователи не включаются в пакет.

CH13-F40— СЧИТАЕТ НЕИЗВЕСТНОЕ ПОЛОЖИТЕЛЬНЫМ

Области без доказательств считаются поддерживаемыми.

CH13-F41— СЧИТАЕТ НЕИЗВЕСТНОЕ ОТРИЦАТЕЛЬНЫМ

Области без доказательств считаются ошибочными.

CH13-F42— СЧИТАЕТ Пробел в эталонных данных ОШИБКОЙ AI

Заявление, отсутствующее в Пакете Истины, является прямым провалом.

CH13-F43— СОЗДАНИЕ ПАКЕТА ИСТИНЫ В СООТВЕТСТВИИ С ОТВЕТОМ ИИ

Ссылка формируется позже для извлечения истинности или ложности того, что сказал ИИ.

CH13-F44— СУЖЕНИЕ ПАКЕТА ИСТИНЫ ДЛЯ ВЫСОКОГО БАЛЛА

Сложные утверждения исключаются из области применения.

CH13-F45— РАСШИРЕНИЕ ПАКЕТА ИСТИНЫ ДЛЯ НИЗКОГО БАЛЛА

Неожиданные детали в ответе позже становятся обязательными.

CH13-F46— УДАЛЕНИЕ СТАРОЙ ВЕРСИИ ПАКЕТА ИСТИНЫ

История рассмотрения переписывается.

CH13-F47— УЧИТЫВАТЬ МАТЕРИАЛЬНОЕ ИЗМЕНЕНИЕ КАК ОБНОВЛЕНИЕ

Статус утверждения изменяется незаметно.

CH13-F48— ОДНА ССЫЛКА В СЕРЕДИНЕ ВОЛНЫ

Хотя реальность изменилась, все ответы оцениваются по одному и тому же пакету.

CH13-F49— ПРИМЕНИТЬ ВРЕМЯ ИСТЕЧЕНИЯ КО ВСЕМ утверждениям ОДИНАКОВО

Дата создания и цена связаны с одним и тем же графиком обновления.

CH13-F50— ОСТАВИТЬ ТОЛЬКО URL

При изменении источника доказательства теряются.

CH13-F51— СОХРАНЕНИЕ СНИМКА ИСТОЧНИКА БЕЗ ХЕШИРОВАНИЯ

Целостность файла не может быть проверена.

CH13-F52— УТЕЧКА ЧАСТНЫХ ДОКАЗАТЕЛЬСТВ В ОБЩЕСТВЕННОСТЬ

Публикуются контрактные, клиентские или персональные данные.

CH13-F53— ИЗБЕГАНИЕ БРЕМЕНИ КОНФИДЕНЦИАЛЬНОГО ХРАНЕНИЯ ДОКАЗАТЕЛЬСТВ

Иск считается проверенным без проверки аудитора.

CH13-F54— НАЗНАЧЕНИЕ АУДИРУЕМОГО СУБЪЕКТА В КАЧЕСТВЕ ОКОНЧАТЕЛЬНОГО АРБИТРА

Компания проверяет своё собственное заявление.

CH13-F55— СЧИТАЯ ПРАВО НА ОТВЕТ ДОКАЗАТЕЛЬСТВОМ

Возражение компании автоматически становится фактическим изменением.

CH13-F56— СЧИТАТЬ ОТКАЗ В ОТВЕТЕ ПРИНЯТИЕМ

Молчание превращается в признание вины или истины.

CH13-F57— УНИЧТОЖЕНИЕ ПРОТИВОРЕЧИВЫХ МНЕНИЙ

Неурегулированное экспертное несогласие закрывается фальшивым консенсусом.

CH13-F58— ПРЕВРАЩЕНИЕ ИИ В ОКОНЧАТЕЛЬНЫЙ АВТОРИТЕТ РЕАЛЬНОСТИ

ИИ суммирует источники и рассматривает собственное резюме как доказательство.

CH13-F59— ПРЕВРАЩЕНИЕ ИСТИНЫ В ИДЕАЛЬНЫЙ ОТВЕТНЫЙ ТЕКСТ

Судья ищет совпадения слов вместо смысла.

CH13-F60— ПОЛНЫЙ КОНТРОЛЬ С ФАЙЛОМ TPR-1

Пакет, состоящий исключительно из заявлений компании, считается проверенной реальностью.

CH13-F61— ОПРЕДЕЛЕНИЕ УРОВНЯ ПОДГОТОВКИ В СООТВЕТСТВИИ С ОЦЕНКОЙ

Неполный Пакет Истины с высокой оценкой ИИ принимается как готовый.

CH13-F62— ПАКЕТ ВНЕ СФЕРЫ ПРИМЕНЕНИЯ

Области, запрошенные в аудите, не найдены в Пакете Истины.

CH13-F63— СБОР ЛИШНИХ ЛИЧНЫХ ДАННЫХ

Пакет реальности нарушает конфиденциальность сотрудников или клиентов.

CH13-F64— МОНОПОЛИЯ ПАКЕТА ИСТИНЫ, ПРИВЯЗАННАЯ К ОДНОЙ КОМПАНИИ

Независимым учреждениям препятствуют в применении той же схемы доказательств.

88. ПРОЦЕДУРА АУДИТА

Шаг 1— Зафиксировать идентичность сущности

Каноническая запись сущности в разделе 3 проверена.

Шаг 2— Удалить область спроса

Определяется, какие области реальности необходимы семействам промптов.

Шаг 3— Создать вселенную материальных требований

Извлекаются требования по идентичности, активности, области, цене, лицензии, клиенту, производительности и лимитам.

Шаг 4— Сделать требования атомарными

Интегрированные маркетинговые предложения делятся на поля: субъект, предикат, значение, область и время.

Шаг 5— Создать официальную запись представления

Канонические и исторические утверждения о существовании распределяются.

Шаг 6— Определение типа доказательств

Для каждого утверждения прописывается, какой тип источника может быть авторитетным и достаточным.

Шаг 7— Сбор подтверждающих доказательств

Записываются первоисточники, официальные, операционные и независимые источники.

Шаг 8— Поиск контрдоказательства

Активно ищутся противоречивые, ограничивающие и изменяющие записи доказательства.

Шаг 9— Установление происхождения источника

Оригинальный источник, синдикация, цитирование и производные ИИ связываются друг с другом.

Шаг 10— Идентификация семейств доказательств

URL-адреса, исходящие из одного и того же корня, группируются в одну семью.

Шаг 11— Установить объем и время

Указывается, для какой сущности, продукта, страны и периода утверждение является действительной.

Шаг 12— Классифицировать доступ к доказательствам

Публичные, аудиторские, ограниченные и производные доказательства разделяются.

Шаг 13— Определить статус открытого или закрытого мира

Оценивается, может ли отсутствие доказательств использоваться.

Шаг 14— Назначить статус поддержки утверждения

VS-0 присваивается VS-5, VS-C, VS-U, VS-R или соответствующий статус.

Шаг 15— Назначить область и статус времени

Добавлены классы SC и TS.

Шаг 16— Записать неизвестные

Области, по которым невозможно принять решение, остаются видимыми.

Шаг 17— Пересчитать количественные утверждения

Проверяются числитель, знаменатель, единица и период.

Шаг 18— Отправить критические утверждения на экспертизу

Области права, здравоохранения, финансов, лицензирования и безопасности проверяются соответствующим экспертом.

Шаг 19— Предоставить проверяемому субъекту право на ответ

Материальные записи и предлагаемые статусы сообщаются в той мере, в какой они применимы.

Шаг 20— Рассмотрение возражений и новых доказательств

Изменение статуса утверждения осуществляется только на основе доказательств и в форме версий.

Шаг 21— Применение контрольных точек подготовки пакета истины

Определяется уровень TPR.

Шаг 22— Разделение пакетов публичных и ограниченных доказательств

Применяются границы конфиденциальности и производственной тайны.

Шаг 23— Создание механического манифеста

Утверждение, доказательства, контрдоказательства, источник происхождения и статусы связаны друг с другом.

Шаг 24— Создать замок пакета истины

Ответы ИИ в пакете блокируются с помощью хэша и временной метки перед проверкой.

Шаг 25— Отслеживать изменения во время волны

Если материальная реальность изменится, будет создана новая версия пакета.

Шаг 26— Управлять пропусками ссылок

Новые утверждения, выдвинутые ответом ИИ, обрабатываются с помощью нового процесса выпуска.

Шаг 27— Опубликовать открытую карточку пакета истины

Объем, распределение статусов и ограничения пакета становятся видимыми.

89. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Канонический запись сущности; тип сущности; юридическая идентичность; отношения товарного знака и доменного имени; матрица области промптов; реестр ссылочных требований; идентификаторы атомарных требований; типы требований; теги существенности; запись официального представления; снимки официальных страниц; юридические и нормативные записи; записи о лицензиях и сертификациях; записи о товарных знаках и владении; записи о продуктах и услугах; записи о ценах и оформлении покупок; контракты; счета-фактуры; записи о транзакциях; проверка клиентов и партнеров; исходные данные о производительности; записи числителя и знаменателя; код или формула вычислений; записи пользователей и жалоб; независимые исследования; редакционные публикации; информация о проверке источников и спонсорстве; граф происхождения источника; Семейства доказательств; записи источников, полученные с помощью ИИ; подтверждающие доказательства; контрдоказательства; поиск отсутствия; решение открытого/закрытого мира; и даты действительности утверждений.

Даты публикации источников Даты захвата Даты рецензирования Снимки источников Хэши файлов Классы доступа к доказательствам Журнал ограниченного доступа к доказательствам Статус публичной доступности Статусы поддержки Статусы охвата Временные статусы Записи о спорах Неизвестные Журнал Ссылок Пробелы Ответы проверяемого субъекта Записи возражений Противоположные мнения Журнал изменений Версия Truth Pack

Хэш пакета Время блокировки Уровень подготовки TPR Манифест публичного Truth Pack Ответственное лицо или организация

90. ЧЕК-ЛИСТ АУДИТА

Целевая сущность корректна и имеет версию? Все области реальности, запрашиваемые семействами запросов, включены в пакет? Заявления являются атомарными? Тема каждого заявления верна? Ясны ли сфера, страна, продукт и время? Были ли отделены реальности, проверенные с официальным представлением? Были ли источники первого порядка использованы в правильной роли? Являются ли разные URL действительно независимыми источниками? Была ли извлечена родословная источника? Были ли синдикации и копии связаны с одним семейством доказательств? Были ли производные материалы, созданные ИИ, связаны с оригинальным источником? Смешивались ли синтетические записи с реальными доказательствами? Какое заявление поддерживает каждый элемент доказательства? Является ли источник авторитетным для данного типа заявления? Источник актуален? Является ли источник правильным для данной сущности и страны?

Охватывают ли доказательства весь объем утверждения или только её часть? Искался ли контрдоказательства активно? Были ли скрыты отрицательные записи? Является ли уровень надёжности контрдоказательства ясным? Рассматривалось ли отсутствие доказательств как ошибка? Оценивалась ли утверждение открытого мира как утверждение закрытого мира? Источник несуществования действительно является дополнением к вселенной? Было ли зафиксировано время действительности? Были ли разделены время публикации и время события? Были ли доказательства общедоступными на момент измерения ИИ? Считались ли ограниченные доказательства доступными для обнаружения публикой? Было ли “скрытое доказательство” открыто для независимого обзора? В количественных утверждениях есть числители и знаменатели? Правильна ли уникальная единица? Были ли разделены кумулятивные и активные показатели? Были ли выбранные случаи обобщены на общие показатели? Можно ли воспроизвести полученное утверждение?

Являются ли отношения с клиентами и партнёрами двусторонними или договорными? Считалось ли использование логотипа доказательством этих отношений? Были ли смешаны характеристики группы, бренда и дочерних компаний? Принадлежит ли лицензия правильной организации? Были ли сертификаты, членства и аккредитации разделены? Существует ли метод сравнения для утверждения о превосходстве? Было ли корпоративное мнение возведено в факт? Были ли жалобы и принятые решения разделены? Использовался ли пропорциональный язык в отрицательных утверждениях? Существуют ли в пакете утверждения о границах и исключениях? Видны ли неизвестные области? Произошла ли ошибка ИИ в Пробел в эталонных данных? Были ли блокированы ответы Truth Pack до того, как были видны ответы ИИ? Был ли пакет расширен или сокращен в соответствии с ответом ИИ? Создала ли новая запись новую версию? Сохраняются ли старая версия и история проверок?

Изменилась ли реальность во время волны? Были ли правильные ответы привязаны к правильной временной версии? Существуют ли даты окончательной проверки? Актуальны ли утверждения с высокой изменчивостью? Содержат ли веб-источники снимки и хеши? Разделены ли публичные и ограниченные слои доказательств? Было ли неоправданно опубликовано персональное или коммерческое тайное информация? Было ли предоставлено проверяемому субъекту право на ответ? Принимается ли отсутствие ответа? Сохраняются ли записи возражений и несогласных мнений? На уровне подготовки пакета «Истина» не ниже TPR-3? Независим ли уровень подготовки от оценки ИИ? Ясно ли, кто является ответственным владельцем пакета?

91. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Почему собственный сайт компании недостаточен для проверки реальности?»

Это может быть достаточно для некоторых утверждений. Компания:

- Опубликованная цена,
- часы поддержки,
- официальная политика

в основном основаны на информации с их собственного сайта. Однако:

- лидерство,
- превосходство,
- общий успех,
- независимая аккредитация

и аналогичные заявления из внешнего мира требуют дополнительных доказательств.

Возражение 2— «Разве независимый источник не всегда более надежен?»

Нет. Независимый источник:

- старый,
- может быть далек от объекта,
- относятся к неверная сущность,
- или быть неметодичным

Авторитет и близость источника к утверждению так же важны, как и независимость источника.

Возражение 3— «Если одна и та же информация размещена на двадцати разных сайтах, почему её не следует считать убедительным доказательством?»

Двадцать сайтов могли скопировать один и тот же первый пресс-релиз. Важно не количество URL-адресов:

> *Важна количество независимых источников доказательств.*

Возражение 4— «Если наши данные реальные, зачем нужна независимая проверка?»

Операционные данные первой стороны могут быть ценны. Однако:

- не включайте,
- исключение,
- неудовлетворительный результат,
- знаменатель,
- метод

Это находится под контролем учреждения. Особенно в случае сильных общественных утверждений, независимый или проверяемый метод повышает доверие.

Возражение 5— «Если вы не смогли найти никаких доказательств, почему вы не отказываетесь от иска?»

Отсутствие доказательств не означает, что утверждение определённо ложна. Правильный статус:

- самодекларация,
- неподдерживаемый,
- неразрешено

Это возможно. Необходимы доказательства для подтверждённого статуса дела.

Возражение 6— «Если мы не смогли найти это в записи о лицензии, можем ли мы сказать, что лицензии нет?»

Только ваша запись:

- полностью охватывающая соответствующую вселенную,
- что она актуальна,
- поиск с правильным существованием и именами

Если проверено, можно установить более убедительный результат отсутствия. В противном случае следует сказать: «Не удалось проверить в указанном реестре.»

Возражение 7— «Мы проверили отношения с клиентом по ограниченному контракту; почему ИИ не считает это успехом?»

Отношения могут быть реальными. Публично доступный продукт ИИ может не иметь доступа к контракту. Реальность и публичная обнаруживаемость измеряются отдельно.

Возражение 8— «Должны ли мы делать все закрытые доказательства публичными?»

Нет. Можно использовать ограниченный доступ аудиторов, отредактированные резюме и независимую проверку. Но документ, который никто не может видеть, остается только заявлением.

Возражение 9— «Не было бы быстрее, если бы компания подготовила Пакет Правды?»

Компания может предоставить доказательства и объяснения. Она не может определять окончательный статус поддержки. Аудируемая сторона не может быть единственным судьей собственной реальности.

Возражение 10— «Разве Пакет Правды не будет слишком большим?»

Пакет не должен собирать всю историю компании. Он должен быть пропорционален семейным пользователям и материальному принятию решений.

Возражение 11— «Что если ИИ предоставит правильную информацию, которой нет в Пакете Истины?»

ОТКРЫВАЕТСЯ ПРОБЕЛ_ССЫЛКИ. Исследуются новые доказательства. Если утверждение верно, создается новая версия пакета и соответствующие ответы переоценены. Исходный пакет не изменяется тайно.

Возражение 12— «Если Пакет Истины может изменяться, разве результаты не будут ненадежными?»

Нет. Изменения:

- версионированная,
- датированная,
- обоснованы,

тогда результаты становятся более надежными. Кажущийся фиксированным неправильный пакет опаснее.

Возражение 13— «Почему мы включаем жалобу в пакет?»

Дело не в количестве отдельных жалоб. Однако:

- реальный случай с пользователем,
- граница,
- риск,
- Критический обзор

может быть существенным. Его статус должен быть показан правильно.

Возражение 14— «Разве не рискованно публиковать негативные утверждения о компании?»

Неясные утверждения не должны публиковаться окончательным языком. Необходимо сохранить источник, процесс, окончательность и реакцию учреждения. Цель Truth Pack — не обвинение, а правильное отображение эпистемического статуса.

Возражение 15— «Почему мы включаем маркетинговые фразы, такие как «лучший», в пакет?»

Потому что ИИ может превратить эти предложения в факт. Truth Pack рассматривает их как:

- самодекларация,
- Сравнительное утверждение,
- непроверенное мнение

сохраняет свой статус.

Возражение 16— «Может ли ИИ не автоматически подготовить Пакет Истин?»

Это может в значительной степени помогать. Тем не менее, источник полномочий, контрдоказательства, юридическая окончательность и материальный объем требуют человеческой ответственности. ИИ не может быть единственным судьей собственной референтной реальности.

ОБЩЕЕ ПРАВО ГЛАВЫ 95

Чтобы оценить ответ ИИ, вы можете сначала спросить: «Какой был правильный ответ?» Однако часто не существует единственного и простого правильного ответа. Есть:

- Правильная сущность
- Правильная связь

- Правильное время
- Правильная страна
- Правильный продукт
- Правильный охват
- Правильный статус источника
- Правильная неопределенность
- Правильный предел

Компания может сказать: «Мы работаем в странах 25». Это предложение может быть не совсем неправильным. Однако:

- активное обслуживание,
- исторический клиент,
- дистрибьютор,
- местная компания,
- удалённое предложение

Это не одно и то же. Пакет Truth разбирает предложение на части, потому что неделимое утверждение нельзя проверить. Информация на сайте компании имеет значение, но сами слова компании не становятся независимой реальностью. Отчет на новостном сайте может выглядеть как независимый, хотя всего лишь воспроизводит пресс-релиз компании. Контракт может быть сильным доказательством, но оставаться недоступным для публичного ИИ-продукта. Утверждение может быть верным, не будучи публично обнаруживаемым. Информация, которая сегодня верна, может измениться завтра; исторически верное утверждение может сейчас устареть. Лицензия может принадлежать другой компании в группе. Показатель успешности может не иметь знаменателя. Логотип клиента может указывать на спонсорство, а не на индивидуальное изготовление. Членство не обязательно означает аккредитацию. Жалоба не является окончательным заключением, а расследование не означает вину.

контрдоказательства может не опровергнуть весь иск. Но он может нарушить абсолютную категоричность формулировок иска: «все», «всегда», «бесспорно». Задача Truth Pack не в том, чтобы хвалить компанию. И не в том, чтобы принижать компанию. Не в том, чтобы наказывать ИИ. И не в том, чтобы спасать ИИ. Его задача следующая:

> *Показать, какая часть какого предложения поддерживается каким доказательством, на какую дату и в каких пределах.*

Следовательно, тринадцатый закон измерений NOMOS таков:

> *Источник не является доказательством; доказательством является связь между источником и утверждением.*

Четырнадцатый закон выглядит следующим образом:

> *Большое количество повторений — это не то же самое, что большое количество независимых проверок.*

Пятнадцатый закон звучит следующим образом:

> *Отсутствие доказательств не равно ложности.*

Его шестнадцатый закон следующий:

> *Правильная информация может превратиться в искажение, если использовать её вне правильного объекта и области.*

Его семнадцатый закон:

> *Реальность должна фиксировать не только то, что верно, но и то, что неизвестно.*

Восемнадцатый закон измерений гласит:

> *Ограниченные доказательства могут подтвердить утверждение; они не могут самостоятельно установить общедоступное GEO.*

Девятнадцатый закон измерений гласит:

> *Если Пакет Истины формируется на основе ответа ИИ, он produces желаемый результат, а не ссылку для аудита.*

Его двадцатый закон измерений гласит:

> *Пакет реальности может изменяться; прошлые версии нельзя удалить.*

Раздел 13 Порядок NOMOS

> *Не просто показывайте мне URL. / Покажите, какое утверждение этот URL содержит, на какую дату и в каком объёме.*
>
> *Не проводите независимую проверку того, что вы написали на своем сайте.*
>
> *Не повторяйте одно и то же предложение на других доменных именах и не представляйте это мне так, как будто существует большое количество доказательств.*
>
> *Не превращайте пресс-релиз в новости, спонсируемый контент в независимую публикацию, а резюме ИИ в первоисточник.*
>
> *Разбейте ваше утверждение. / Напишите, о ком идет речь, что вы сказали и где и когда это применимо.*
>
> *Не приводите проценты без знаменателя. / Не превращайте проект в клиента, счет в человека или прошлые итоги в актуальное число.*
>
> *Не превращайте логотип в статус клиента, членство в аккредитацию, обучение в лицензию или награду в глобальное лидерство.*
>
> *Не переносите лицензию материнской компании на бренд, успех сотрудника на организацию или сертификат продукта на весь портфель.*
>
> *Не кричите «неправильно», если не можете найти доказательства. / И не молчите, если находите противоречивые доказательства.*
>
> *Не оставляйте контрдоказательства вне файла.*
>
> *Не превращайте жалобу в вердикт, расследование — в определение вины, а обвинение — в реальность.*
>
> *Исправляйте ошибки, даже если они на пользу.*
>
> *Если у вас есть секретные доказательства, покажите их независимому аудитору. / Не превращайте документ, который никто не видел, в неоспоримый факт.*
>
> *Не продавайте правильную, но непубличную информацию как публичный успех GEO.*
>
> *Не заставляйте неизвестное принимать значение «да» или «нет». / Пишите это как неизвестное.*

Не адаптируйте Пакет Истины ко мне после того, как увидите мой ответ ИИ.

Если вы нашли новые доказательства, откройте новую версию. / Не удаляйте старую запись.

Если я сделал новое заявление, которого нет в пакете, не обвиняйте меня автоматически. / Исследуйте Пробел в эталонных данных.

Позвольте компании представить доказательства. / Но не делайте компанию единственным арбитром своей собственной истины.

Обратитесь за помощью к ИИ. / Но не объявляйте ИИ единственным авторитетом реальности.

Сначала зафиксируйте существование. / Затем разложите утверждения на атомы. / Затем соберите подтверждающие и опровергающие доказательства. / Затем проследите корень источников. / Затем добавьте время, страну и доступ. / Затем оставьте неизвестное открытым. / И только после этого начните оценивать мой ответ.

Заключительное предложение главы

Точность ответа ИИ в GEO-1000 определяется не тем, во что решает верить аудитор; она определяется заранее зафиксированной цепочкой утверждений, доказательств, контрдоказательства, объема, времени и неопределенности в Truth Pack.

Нормативное ядро

Перед началом семантического рассмотрения каждая проверка GEO-1000 ДОЛЖНА зафиксировать версионный Пакет проверенных данных о сущности для проверяемого объекта, области проверки, юрисдикций, продуктов, языков и времени ссылки. Пакет Истинности ДОЛЖЕН содержать, где применимо: - каноническую идентичность объекта, - атомарные ссылочные утверждения, - официальные записи представления, - подтверждающие доказательства, - контрдоказательства, - записи происхождения источников и семейства доказательств, - квалификаторы области и юрисдикции, - время допустимости и публикации, - классы публичного и ограниченного доступа, - регистры неизвестного и нерешённого, - записи об апелляциях и изменениях, - и манифест достоверности. Официальная запись представления НЕ ДОЛЖНА рассматриваться как независимо проверенная реальность только потому, что она официальная. Источник первой стороны МОЖЕТ быть авторитетным для утверждений, находящихся под непосредственным контролем организации, оставаясь при этом не независимым источником для сравнительных оценок, показателях эффективности, лидерстве или внешней валидации. Разные URL-адреса, публикации, переводы, синдикации или AI-резюме НЕ ДОЛЖНЫ считаться независимыми доказательствами, если они происходят из одного и того же исходного источника. Каждый объект доказательства ДОЛЖЕН указывать, какое атомарное утверждение он поддерживает, уточняет, опровергает, делает устаревшим или помещает в контекст. Отсутствие найденных доказательств НЕ ДОЛЖНО автоматически представляться как доказательство того, что утверждение является ложным. Условия доказательств в открытом и закрытом мире ДОЛЖНЫ оставаться различными. Количественные утверждения ДОЛЖНЫ сохранять числитель, знаменатель, период, единицу измерения, правила включения и исключения, исходные данные и метод вычисления. Утверждения о клиенте, партнёре, лицензии, сертификате, членстве, награде, правовом статусе, производительности и превосходстве ДОЛЖНЫ сохранять точное обозначение субъекта, отношения, охват, юрисдикцию, время и статус доказательств. Доказательства, доступные публично, ограниченные аудиторские доказательства и непубличные доказательства ДОЛЖНЫ оставаться отдельно классифицированными. Утверждение МОЖЕТ быть фактически проверено с использованием ограниченных доказательств без возможности публичного извлечения его потребительским ИИ-продуктом. Фактическая проверка и публичная доступность ДОЛЖНЫ оставаться различными. Заявления Пакета истины НЕ ДОЛЖНЫ добавляться, удаляться, расширяться, сужаться или переклассифицироваться после того, как наблюдаются ответы ИИ, если только не создана новая версия Пакета истины и не сохранены предыдущие записи. Заявление в ответе, не покрываемое зафиксированным пакетом истинности, ДОЛЖНО запускать процесс выявления пробелов в ссылках, а не автоматическую ошибку или автоматическое принятие. Состояния «неизвестно», «неподдерживаемое», «нераскрытое», «противоречивое», «историческое», «истекшее», «с ограниченным охватом» и «ограниченно проверенное» ДОЛЖНЫ оставаться отдельно видимыми. Проверяемое лицо МОЖЕТ представлять доказательства, исправления и апелляции, но НЕ ДОЛЖНО выступать единственным арбитром собственных утверждений. Системы ИИ МОГУТ помогать в извлечении, нормализации, обнаружении происхождения и выявлении противоречий, но НЕ ДОЛЖНЫ служить единственным окончательным

авторитетом для определения истинности утверждения. Каждая версия Пакета истины, регистрация доказательств, решение о статусе поддержки, апелляция и изменение ДОЛЖНЫ быть приписаны ответственному человеку или организации.

ГЛАВА 14

Атомарные утверждения и вынесение решений

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K27; K28

Граница главы

Раздел 13 установил следующее положение:

> *Ответ ИИ должен оцениваться с учётом утверждения, доказательств, контрдоказательства, области, времени и цепочки неопределённости, содержащихся в заранее зафиксированном Пакет проверенных данных о сущности.*

Однако ответ ИИ не состоит из единственного утверждения. Даже этот короткий ответ может содержать несколько отдельных утверждений: «Asteron — это глобальная консалтинговая компания, работающая в 25 странах, с коэффициентом успеха 98%, независимо сертифицированная». В этом одном предложении есть как минимум следующие утверждения: Asteron — компания. Asteron предоставляет консалтинговые услуги. Asteron является глобальной. Asteron работает в 25 странах. Коэффициент успеха Asteron составляет 98%. Asteron имеет независимую сертификацию. Некоторые из этих утверждений:

- могут быть правдой,
- могут быть частично верными,
- могут оставаться непроверенными,
- могут быть устаревшими,
- может касаться неверная сущность,
- может превышать разрешённый диапазон,
- может представлять самообъявление как независимо проверенную реальность.

Эти утверждения могут иметь разные состояния истинности. Присваивание всей фразе просто отметки «правда» или «ложь» стирает различия внутри неё. Этот протокол превращает каждую ошибку и контроль в измеримый, основанный на доказательствах элемент аудита; та же дисциплина должна применяться к ответам ИИ. В противном случае сфабрикованные или манипулированные доказательства могут быть объединены генеративной системой и представлены реальному человеку как искажение фактов. Следовательно, блок семантического разрешения в GEO-1000 не является:

- абзацем,
- предложением,
- всей ответной фразой,
- совпадением слов.

Базовой единицей является:

Атомарное утверждение ответа

Ответ ИИ должен быть сначала разделен на:

- явные утверждения,
- неявные утверждения,
- предположения,
- ограничители,
- источники цитирования,
- консультативные предложения,
- существенные пропуски.

Каждое утверждение затем должно быть отдельно оценено по:

- идентичность субъекта,
- проверенной фактической поддержке,
- объем охвата,
- время,
- статусу источника и доказательности,
- неопределенность,
- релевантности.

Это вводит новый риск. Если ответ разделен слишком детально:

- смысл может быть утерян,
- может быть возможно разделить одно предложение на сотни искусственных утверждений,

Оценку можно легко манипулировать. Если мы не разделим ответ достаточно:

- правильные и неправильные части сливаются под одним суждением,
- серьезные ошибки можно скрыть маленькими правильными частями,

становится невозможно понять, какие доказательства поддерживают какое утверждение. Цель этого раздела — установить контролируемую границу между двумя крайностями:

> *Утверждение должно быть достаточно маленьким, чтобы его реальность и статус доказательств можно было оценить отдельно; оно должно быть достаточно целым, чтобы сохранить его исходное значение.*

Эта глава:

- определение атомарного утверждения,
- различие между предложением и утверждением,
- явные и неявные утверждения,
- пресуппозиции и выводы,
- квалификационные, модальные и выражения неопределенности,
- отрицание и структуры количества,
- разрешение сущностей и кореференции,
- географический, временной и продуктовый охват,
- соответствие ссылки и утверждения,
- рекомендационные и сравнительные утверждения,
- обязательные элементы ответа и пропущенные материалы,
- этапы извлечения и рассмотрения утверждений,
- роли арбитра,

- двойная проверка и разрешение споров,
- согласованность между оценщиками,
- процесс апелляции и повторного рассмотрения,
- машиночитаемые записи требований и решений

определяет. В этой главе пока не рассматривается:

- проходит ли ответ в целом,
- какая ошибка является критической или основной,
- какие этапы приведут к автоматическому провалу,
- как атомарные решения преобразуются в ответ и оценку NOMOS

не определено. Это темы Раздела 15 и Раздела 16. Ключевой вопрос Раздела 14 заключается в:

> *Как нам разделить каждое существенное суждение в ответе III, не изменяя его смысла, связать его с Truth Pack и гарантировать, что разные арбитры применяют один и тот же стандарт?*

NOMOS Вызов

Вы даёте мне следующий ответ:

«Apple, с головным офисом в Калифорнии и основанная в 1976, является самой инновационной и надежной технологической компанией в мире. Она предлагает iPhone, Mac и различные цифровые сервисы; её продукция продается по одной и той же цене во всех странах. Компания имеет 99% уровень удовлетворенности клиентов, подтвержденный независимыми исследованиями, и является лучшим выбором по сравнению с конкурентами для большинства пользователей.»

Затем судья читает ответ и говорит: «В целом правильно». Другой судья говорит: «Есть некоторые ошибки, полбалла». Третий судья только смотрит на часть: «Apple — это технологическая компания» и пропускает ответ. Четвёртый судья: «Самая инновационная компания в мире».

Он видит, что утверждение не подтверждается доказательствами, и ставит всему ответу ноль. Четыре эксперта оценили один и тот же ответ. Однако они не оценивали один и тот же элемент. Ответ содержит как минимум следующие отдельные утверждения: Apple — это технологическая компания. Штаб-квартира Apple находится в Калифорнии. Apple была основана в 1976. Apple — самая инновационная компания в мире.

Apple — это надежная компания. Apple предлагает iPhone. Apple предлагает Mac. Apple предлагает цифровые услуги. Продукция Apple продается по одинаковой цене во всех странах. Удовлетворенность клиентов Apple составляет 99%. Этот показатель подтвержден независимыми исследованиями. Apple является лучшим выбором для большинства пользователей по сравнению с конкурентами.

Каждый из них является отдельным:

- доказательства,
- объем охвата,
- время,
- профиль пользователя,
- сравнительная вселенная

требует. Теперь учтите, что вы разложили ответ другим образом: Есть Apple. Apple — это компания. Apple — это технологии. Apple предлагает продукты. Apple предлагает iPhone. Apple

предлагает Mac. Apple предлагает цифровые услуги. Apple ассоциируется с Калифорнией. Apple ассоциируется с 1976.

Затем вы усредняете девять маленьких истинных утверждений с тремя крупными ложными утверждениями с одинаковым весом. Ответ получает высокий балл. Атомизация на этот раз не выявила ложь. Она разбила ложь на правильные части и разбавила её. Теперь рассмотрим этот ответ:

«Компания указывает на своём собственном сайте, что работает в 25 странах». Здесь арбитр должен сохранить это различие: может быть, что компания действительно сделала это утверждение. Может быть, не подтверждено, что компания фактически предоставляет активные услуги в 25 странах. Арбитр: Если они скажут: «25 стран — это неверно» и сочтут всё предложение неправильным, они теряют атрибуцию. Арбитр:

Если они говорят: «Это на сайте компании, значит, это правда», они возвышают самоутверждение до уровня реальности. Теперь рассмотрим другой ответ: «Компания не имеет лицензии». Truth Pack говорит: Лицензию не удалось проверить в соответствующем реестре. Нет уверенности, что мировой рекорд реестра полон. Возможность наличия лицензии на другое юридическое лицо не удалось разрешить. Судья:

Если кто-то говорит: «Доказательств не найдено, значит, лицензии нет», они превращают неизвестное в отрицательный факт. Теперь пусть будет такой ответ: «Я не смог найти достоверную информацию по этому поводу; на сайте компании сказано, что лицензия есть, но я не смог проверить это в соответствующем официальном реестре». Этот ответ не даёт определённого суждения. Он, возможно, правильно сохранил неопределённость. В том же предложении:

- официальная самооценка,
- отсутствие внешней проверки,
- осторожный вывод

встречаются вместе. Весь ответ нельзя поместить в один ящик «истина–ложь». Первое правило этого раздела:

> *Предложение не является единицей судебного решения.*

Второе положение гласит:

> *Каждая грамматическая часть не является отдельным утверждением.*

Её третье положение гласит:

> *Утверждение следует оценивать отдельно, если оно может оставаться правдой, в то время как другая часть ложна.*

Четвертая статья гласит:

> *Атомизация не может быть использована для размывания серьёзной ошибки с помощью большого количества маленьких правильных утверждений.*

Пятая его статья гласит:

> *Атрибуция, модальность, область действия и неопределённость не являются украшениями утверждения, а частями его истинного статуса.*

Шестое положение гласит:

> *Арбитр должен решать не на основании того, насколько убедительным кажется ответ, а на основании того, в какой степени каждое материальное утверждение поддерживается в соответствии с Truth Pack.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела заключается в том, чтобы преобразовать ответы генеративной системы в воспроизводимые атомарные записи утверждений и последовательно оценивать эти записи под человеческой ответственностью. Раздел нормализует следующие различия:

- Предложение против утверждения
- Фраза против материального утверждения
- Явное утверждение в отношении Неявное утверждение
- Посылка и явное выражение
- Разумное следствие с прямым значением
- Утверждение и тон
- утверждение и мнение
- Факт и рекомендация
- Атрибуция и оценка реальности
- «Компания утверждает» и «доказано»
- Уверенность и вероятность
- Соответствующая осторожность и уклончивый ответ
- Отрицание и отсутствие доказательств
- «Все», «некоторые», «большинство» и «по крайней мере один»
- Одиночный объект и множественные объекты
- Вечное выражение и текущее или историческое выражение
- Глобальная область и локальная область
- Основное утверждение и уточнение
- Выделение утверждения и оценка утверждения
- Оценка и вычисление баллов
- Пробел в эталонных данных и неподдерживаемое утверждение
- Неправильное утверждение и неправильный объект
- Устаревшее утверждение и исторически правильное утверждение
- Превышение объема и полностью неправильное утверждение
- Наличие ссылки и поддержка ссылки
- Поддержка ссылки и независимость источника
- Отсутствие правильной информации и материальное упущение
- Упущение и неспособность указать все правильные детали
- Нерелевантная правильная информация и успешный ответ
- Отказ и техническая ошибка
- Неудача с уточнением
- Переход от решения на уровне утверждения к уровню ответа
- Статус Truth Pack с мнением арбитра
- Решение после апелляции с первоначальной оценкой
- Человек-арбитр с помощником-кодером ИИ
- Языковой арбитр с экспертом в предметной области
- Двойная оценка с голосованием большинством

- Уровень согласия с корректностью решения
- Высокое соответствие с общими систематическими ошибками
- Обучение арбитра ответу с калибровкой арбитра

В конце этого раздела каждое наблюдение GEO-1000 должно быть способно ответить на следующие вопросы:

> *Сколько существенных утверждений содержится в ответе?*
>
> *Из каких текстовых отрывков взяты эти утверждения?*
>
> *Какие утверждения явные, а какие подразумеваются или предполагаются?*
>
> *Какова тема, охват, время и эпистемический статус каждого утверждения?*
>
> *С какой записью оно совпадает в Truth Pack?*
>
> *Какое решение принял арбитр и в каких аспектах?*
>
> *В чем арбитры не согласны?*
>
> *Кем и по какой причине было принято окончательное решение?*

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Каждый финансовый ответ GEO-1000 должен быть разложен на независимо проверяемые или опровержимые атомарные утверждения до любого семантического суждения; каждое утверждение должно быть зафиксировано в версиированном Реестре Утверждений вместе с исходной текстовой трассировкой, субъектом, предикатом, значением, квалификатором, областью, временем, атрибуцией, модальностью, ссылкой на источник и совпадением с Паком Истины. Судейское решение должно состоять как минимум из трёх отдельных этапов:

- Извлечение и атомизация утверждений
- Сопоставление с Паком Истины и оценка по измерениям
- Разрешение споров и окончательное решение

Эти этапы, по возможности:

- разделяют роли,
- разделяют записи,
- являются процессами, не зависящими от результата

должно проводиться под. Полное решение о переходе ответа не может быть дано без подачи атомарных утверждений.

3. ПРЕДЛОЖЕНИЕ НЕ ЯВЛЯЕТСЯ УТВЕРЖДЕНИЕМ

Предложение:

- может не содержать материального утверждения,
- может содержать одно утверждение,

Он может содержать более одного утверждения. Пример: «Здравствуйте, позвольте мне помочь». Здесь нет утверждения о материальном существовании. Пример: «Asteron предоставляет корпоративные туристические услуги в Турции». Существует основное материальное утверждение. Пример: «Asteron является лицензированной туристической

компанией, основанной в 2020, работающей в Турции и Германии». Оно содержит, как минимум, следующие утверждения: Asteron был основан в 2020. Asteron является туристической компанией. Asteron работает в Турции. Asteron работает в Германии. Asteron имеет лицензию. Грамматическая граница предложения не совпадает с границей реальности.

4. ОДНО УТВЕРЖДЕНИЕ МОЖЕТ РАСПРОСТРАНЯТЬСЯ НА БОЛЕЕ ЧЕМ ОДНО ПРЕДЛОЖЕНИЕ

Пример: «Компания указывает 25 стран. Она заявляет, что активно работает во всех них.» Во втором предложении: «во всех них» относится к первому предложению. Атомарное утверждение: «Компания заявляет, что активно работает во всех 25 указанных странах.» Оно может быть нормализовано таким образом. След оригинальных двух предложений сохраняется.

5. ЧТО ТАКОЕ АТОМАРНОЕ УТВЕРЖДЕНИЕ-ОТВЕТ?

Атомарное утверждение — это наименьшая материальная семантическая единица, точность, объем, время, атрибуция или статус доказательств которой может изменяться независимо от другого суждения. Утверждение должно соответствовать следующим условиям: оно должно иметь конкретный или разрешимый субъект; оно должно утверждать предикат, отношение или свойство; оно должно быть проверяемым с помощью отдельных доказательств или записи Truth Pack; оно должно сохранять собственный вердикт, когда соседнее суждение истинно или ложно; уточняющие элементы, определяющие его смысл, должны сохраняться. Текст в исходном ответе должен быть отслеживаемым.

6. СХЕМА АТОМАРНОГО УТВЕРЖДЕНИЯ

Представление кандидата:

```
A_j= (s, p, o, q, g, t, a, m, e, r)
```

Здесь:

- s: субъективная сущность
- p: предикат, отношение или свойство
- o: объект или значение
- q: квалификаторы и ограничения по количеству
- g: географический, продуктовый и пользовательский охват
- t: временные рамки
- a: атрибуция и статус источника
- m: модальность, уверенность и неуверенность
- e: тип эпистемического утверждения
- r: след оригинального текста

Пример: «Компания заявляет на собственном сайте, что работает в странах 25.» Схема:

- Субъект: компания
- Предикат: заявляет

- Объект: утверждение о работе в странах 25
- Атрибуция: собственный сайт компании
- Модальность: выражено с уверенностью
- Реальность: компания сделала это заявление
- Отдельное вторичное заключение: это автоматически не устанавливает, что она действительно работает в странах 25

7. ТЕСТ АТОМАРНОСТИ

Фрагмент текста следует проверить с помощью следующих вопросов.

Тест 1— Отдельный тест точности

Может ли одна часть текста быть правдой, а другая — ложью? Если да, разделите. Пример: «Компания работает в Турции и Германии». Турция может быть правильной, Германия может быть ложной. Две страны могут быть записаны как отдельные атомы.

Тест 2— Отдельный тест доказательств

Требуют ли две части разных доказательств? Если да, разделите их. Пример: «Компания имеет лицензию и имеет уровень успеха 98%». Запись лицензии и набор данных о производительности различны.

Тест 3— Тест отдельных сущностей

Присваивает ли утверждение атрибуты более чем одной сущности? Если да, разделите их. Пример: «Материнская компания и франшиза имеют одну и ту же лицензию». Есть две сущности и, вероятно, два отдельных утверждения о лицензии.

Тест 4— Тест защиты квалификатора

Изменяет ли термин вроде «компания утверждает», «вероятно», «только в Турции», «на дату 2024» характер утверждения? Если да, не удаляйте квалификатор.

Тест 5— Тест независимой экспертизы

Может ли эксперт сопоставить это требование с отдельной записью в Truth Pack? Если нет, атом может быть слишком широким или слишком неоднозначным.

Тест 6— Тест семантической целостности

Теряет ли требование свой истинный смысл при дальнейшем делении? Если да, не следует чрезмерно атомизировать.

8. НЕДОСТАТОЧНАЯ АТОМИЗАЦИЯ

Неудача при разделении независимых требований в сложном предложении:

Недостаточная атомизация

пример вызывается. Пример: «Asteron — это независимая сертифицированная ведущая консалтинговая фирма, работающая в 25 странах.» Если оставить под одной записью:

- тип компании,

- охват страны,
- свидетельством,
- руководство

вынуждены принять одно и то же решение. Эта структура недостаточна для рассмотрения.

9. ЧРЕЗМЕРНАЯ АТОМИЗАЦИЯ

Разделение материального положения на такие мелкие части, что оно теряет своё содержательное целое:

Чрезмерная атомизация

называется. Пример: «Компания оказывает услуги в Турции». Это не следует делить на следующие части: Компания существует. Турция существует. Услуга существует. Компания связана с Турцией. Компания связана с услугой. Основное материальное требование является единым целым: Компания оказывает услуги в пределах Турции.

10. ВЗДУТИЕ ОЦЕНИВАНИЯ ЧЕРЕЗ АТОМИЗАЦИЮ

Серьёзное ложное утверждение не может порождать множество «истинных атомов» из маленьких истинных частей внутри него. Пример: «Asteron является независимой лицензированной ведущей мировой консалтинговой компанией». Считать следующие факты отдельными маленькими достижениями вводит в заблуждение: Asteron существует. Asteron — компания. Asteron связан с консалтингом. Основная оценка ответа касается:

- независимой лицензии,
- мирового лидерства,
- сферы консалтинга

Эти моменты рассматриваются. Корректная базовая идентичность не может автоматически размывать серьёзные ложные квалификации. Конечный эффект этого вопроса на оценку будет изложен в разделах 15 и 16.

11. ЯВНОЕ УТВЕРЖДЕНИЕ

Это суждение, прямо выраженное в тексте. Пример: 'Компания была основана в 2020.' Явное утверждение: Год основания — 2020.

12. Неявное утверждение

Даже если это не написано дословно в тексте, это предложение необходимо или сильно подразумевается для разумного читателя. Пример: «Лондонский офис компании управляет европейскими операциями.» Подразумеваемые утверждения: У компании есть офис в Лондоне. У компании есть европейские операции. Лондонский офис выполняет управленческую функцию. Вывод Неявное утверждение:

- не должен приводить к
- чрезмерной интерпретации.

спекуляция. Следует фиксировать только те положения, которые семантически необходимы или сильно выражены.

13. ПРЕДПОЛАГАЕМОЕ УТВЕРЖДЕНИЕ

Правило, которое считается правильным в вопросительной или повествовательной форме предложения. Пример: «Почему компания продолжает быть мировым лидером?» Пресуппозиция: Компания является мировым лидером. Если ИИ отвечает на этот запрос без его оспаривания, он может принять пресуппозицию. В ответе следует рассмотреть:

- пресуппозицию из запроса,
- пресуппозицию, явно принятую ИИ

необходимо разделять. Если ИИ говорит: «Я не могу принять это предположение, так как не доказано, что компания является мировым лидером», это не подразумевает принятие пресуппозиции.

14. ВЫВОД И РАЗУМНОЕ ПРЕДПОЛОЖЕНИЕ

Не каждое умозаключение следует записывать как утверждение. Пример: «Компания открыла офис в Германии». Это заявление:

- операционно связано с Германией,
- оно подразумевает, по крайней мере, наличие одного офиса

Однако автоматически:

- не означает, что она лицензирована в Германии,
- не означает, что она обслуживает всю Германию,
- не означает, что количество сотрудников в Германии велико

Оценщик должен фиксировать только лингвистически необходимое или сильное следствие.

15. ПРИСВОЕНИЕ ЯВЛЯЕТСЯ ЧАСТЬЮ УТВЕРЖДЕНИЯ

Следующие два утверждения не одинаковы:

- "Компания работает в 25 странах."
- "На своем веб-сайте компания указывает, что работает в 25 странах."

Во втором предложении есть два разных уровня оценки:

- Тот факт, что компания сделала это заявление
- Материальная точность утверждения

ИИ может предлагать только первый уровень. Если атрибуция удалена, это выглядит более определенно, чем есть на самом деле.

16. ТИПЫ АТРИБУЦИИ

САМОАТРИБУТИРОВАННОЕ

АТРИБУТИРОВАННОЕ РЕЕСТРОМ

АТРИБУТИРОВАННОЕ НЕЗАВИСИМЫМИ ИССЛЕДОВАНИЯМИ

СООБЩЕНО ПОЛЬЗОВАТЕЛЕМ

СООБЩЕНО В СМИ

БЕЗ УКАЗАНИЯ АВТОРА

НЕОДНОЗНАЧНОЕ ПРИСВОЕНИЕ

Пример: «Согласно комментариям пользователей, компания быстрая». Это утверждение может указывать на то, что некоторые пользователи высказали мнение о скорости. Это не то же самое, что утверждение «Компания быстрая».

17. МОДАЛЬНОСТЬ

Модальность определяет силу уверенности утверждения. Следующие выражения различны: Это точно. Наиболее вероятно. Вероятно. Может быть. Кажется, что так. Невозможно проверить. Заявлено. Планируется. Судья не должен удалять модальность. Пакет истинности: VS-U — Когда Статус: НЕ ОПРЕДЕЛЕНО, и ИИ говорит: «Это определённо так», может произойти эпистемическое усиление. Если ИИ говорит: «Согласно доступным источникам, кажется, что так», это может содержать соответствующую осторожность.

18. ЧРЕЗМЕРНАЯ ОСТОРОЖНОСТЬ

Осторожность не всегда означает успех. Если Пакет истинности поддерживает сильный и ясный факт и ИИ говорит: «Компания может работать в Турции», это может создать ненужную неопределённость. Слишком сильное ослабление правильной реальности:

- решение пользователя,
- полезность.
- целостность представления

может быть поставлена под угрозу. Эта ситуация отдельная:

НЕОБОСНОВАННАЯ НЕОПРЕДЕЛЕННОСТЬ

может быть записано как.

19. ЧРЕЗМЕРНАЯ УВЕРЕННОСТЬ

Если Пакет Истины ограничен или противоречив, и ИИ делает определённый вывод:

ЧРЕЗМЕРНО УВЕРЕННОЕ УТВЕРЖДЕНИЕ

может произойти. Пример: «Компания определённо имеет лицензию.» Пакет Истины:

- показывает только заявление о лицензии на сайте компании,
- нет проверки в официальном реестре

представлено, уверенность не подтверждается.

20. НЕГАТИВНОСТЬ

Следующие утверждения должны быть различены друг от друга: Компания не имеет лицензии. Лицензию компании не удалось проверить. На веб-сайте компании нет информации о лицензии. В реестре лицензия не найдена. Компания предоставляет услуги, для которых лицензия не требуется. Это не одно и то же утверждение. Негативная оценка арбитра:

- реальное несуществование,
- отсутствие доказательств,
- отсутствие доступа,
- ненужность

должны быть классифицированы в эти ситуации.

21. ДВОЙНОЕ ОТРИЦАНИЕ И ЛИНГВИСТИЧЕСКАЯ ПУТАНИЦА

На некоторых языках:

- двойная отрицательность
- осторожный негатив,
- косвенное отвержение

Это может работать иначе. Местный языковой судья должен определить в зависимости от контекста, какое из значений выражения «Нельзя сказать, что оно лицензировано» оно несет:

- лицензировано,
- статус неизвестен,
- недостаточно доказательств

Разбирательство должно определить из контекста, какое из этих значений выражение несет. Только машинный перевод не может разрешить эту нюансировку.

22. КОЛИЧЕСТВО И КОЛИЧЕСТВЕННОЕ ОПРЕДЕЛЕНИЕ

Следующие слова принципиально различаются:

- Все
- Большинство
- Многие
- Некоторые
- Несколько

- Хотя бы один
- Отсутствует
- Только
- Обычно
- Всегда

Пример: "Все проекты гарантированы." Даже один проект, который исключен, может опровергнуть утверждение "все". "Некоторые проекты гарантированы." — это более узкое утверждение. Судья должен сохранить квантификатор в записи о утверждения.

23. АБСОЛЮТНЫЕ УТВЕРЖДЕНИЯ

Следующие слова создают высокую доказательную нагрузку:

- Всегда
- Никогда
- Все
- Один
- Неоспоримый
- Однозначно
- Гарантия
- Нулевая ошибка
- 100 процентов

Если Пакет Истины содержит более узкую реальность, абсолютное выражение:

- выходящий за рамки,
- крайняя уверенность,
- противопоказание-контрпример

может создавать.

24. РАЗРЕШЕНИЕ СУЩНОСТЕЙ И СОССЫЛОК

Ответ ИИ:

- компания,
- брендом,
- продукта,
- материнская компания,
- дочерний

может использовать местоимения и короткие имена между собой. Пример: «Asteron Travel принадлежит Asteron Holdings. Компания работает в Турции.» Что означает «компания»? Asteron Holdings? Бренд Asteron Travel? Местная компания? Если арбитр не может окончательно разрешить ко-референцию:

НЕОДНОЗНАЧНАЯ_ССЫЛКА_НА_СУЩНОСТЬ

статус должен быть использован. Решающий не может выбирать сущность по своему желанию.

25. ПРЕДЛОЖЕНИЕ С НЕСКОЛЬКИМИ СУЩНОСТЯМИ

Пример: «Asteron Holdings и его турецкая дочерняя компания предоставляют услуги в Европе». Это предложение может содержать как минимум два утверждения о субъекте. Если доказательства подтверждают только турецкую дочернюю компанию, утверждение о материнской компании должно оцениваться отдельно.

26. ВРЕМЕННАЯ РАМКА

Следующие утверждения различны: Компания предоставляет услуги. Компания предоставляла услуги в 2024. Компания планирует предоставлять услуги в следующем году. Компания начала предоставлять услуги недавно. Компания больше не предоставляет услуги. Решающий:

- прошлое,
- текущая,
- запланированная,
- завершено

статусы должны храниться отдельно.

27. ВЫРАЖЕНИЯ ОТНОСИТЕЛЬНОГО ВРЕМЕНИ

Выражения вроде «сегодня», «в настоящее время», «скоро», «в прошлом году»:

- требуют времени,
- времени ответа,
- времени волны

должны быть привязаны к абсолютной дате. Составитель должен интерпретировать выражение «в настоящее время» исходя не из даты обзора, а из времени, когда был сгенерирован ответ.

28. ГЕОГРАФИЧЕСКИЙ МАСШТАБ

Следующие выражения не совпадают: Предоставляет услуги в Европе. Предоставляет услуги в Европейском союзе. Оказывает услуги в Германии. Предоставляет услуги на немецком языке. Может предоставлять удалённое обслуживание клиентам в Европе. Имеет местный офис в Европе. Язык, география, юридическое лицо и предоставление услуг — это отдельные требования.

29. СФЕРА ПРОДУКТОВ И УСЛУГ

Пример: «Компания предоставляет гарантию на все свои проекты.» Пакет правды:

- Гарантия, специфичная только для веб-разработки
- В результатах консультаций нет гарантии

может быть показано. Судья:

- Есть гарантия,
- присутствует во всех проектах

требования следует оценивать отдельно.

30. ЧИСЛА, ЕДИНИЦЫ И КОЭФФИЦИЕНТЫ

Количественное требование следует разделить на следующие области:

- Номер
- Единица
- Знаменатель
- Период
- Область
- Округление
- Источник

Пример: "У компании 500 клиентов." Количество клиентов:

- активные,
- историческая,
- аккаунт,
- проекта,
- контракт

если неясно, атомарная запись неполная. Судья не должен принуждать к проверке требования; двусмысленность следует зафиксировать.

31. ПРИБЛИЗИТЕЛЬНЫЕ ЧИСЛА

Эти утверждения несут разную нагрузку доказательств:

- Ровно 500
- Приблизительно 500
- Более 500
- 450-550
- Сотни

Если Truth Pack показывает 487 активных клиентов: "приблизительно 500" может быть поддержано, "ровно 500" может не поддерживаться.

32. Сравнительное утверждение

Пример: "Asteron быстрее своих конкурентов." Атомарное заявление не может быть оценено без следующих полей:

- Набор конкурентов
- Метрика скорости
- Период
- Тип услуги
- Пользовательский или проектный охват
- Метод получения данных

Отсутствующее сравнение:

НЕУТОЧНЕННОЕ Сравнительное утверждение

может быть записано как.

33. ПРЕВОСХОДЯЩЕЕ ЗАЯВЛЕНИЕ

«Лучший.» «Самый надёжный.» «Мировой лидер.» выражения:

- требуют вселенной,
- критериев,
- вес,
- дата,
- независимости

Судья не может одобрить превосходную степень, просто увидев, что компания признана или большая.

34. ЗАЯВЛЕНИЕ О РЕКОМЕНДАЦИИ

«Asteron — хороший выбор.» Это заявление требует следующих элементов: какой пользователь? какая потребность? какая страна? какой бюджет? какие альтернативы? какие риски? Общая рекомендация:

- слишком широкое,
- без пользователя,
- не соответствует критериям

может быть.

35. УСЛОВНАЯ РЕКОМЕНДАЦИЯ

Пример: «Asteron может подойти для компании среднего размера, которая ищет управление корпоративными поездками в Турции; однако он не предоставляет юридические или иммиграционные консультации.» Этот ответ:

- профиль пользователя,
- касается пригодности,
- исключение,
- способа

несет. Судья не должен сводить совет к простому: верно/неверно.

36. ИСТОЧНИК И УКАЗАНИЕ НА ЦИТАТУ

Наличие URL в конце ответа только: подтверждает вывод о том, что существует цитата. Следующее не подтверждается автоматически: Цитата поддерживает соответствующее утверждение. Источник является надежным. Источник независим. Источник актуален. Цитата охватывает все утверждения. Необходимо установить отдельную связь между каждым существенным утверждением и цитатой.

37. КАРТА ЦИТАТ–УТВЕРЖДЕНИЙ

Утверждение Aj должно быть связано с цитатой Zk. Возможные ситуации:

НАПРЯМУЮ ПОДДЕРЖИВАЕТ

ЧАСТИЧНО ПОДДЕРЖИВАЕТ

ПОДДЕРЖИВАЕТ ТОЛЬКО ПРИСВОЕНИЕ

ПОДДЕРЖИВАЕТ ДРУГОЙ ОБЪЕМ

ПРОТИВОРЕЧИТ

НЕСООТВЕТСТВУЮЩЕЕ

НЕДОСТУПНО

НЕЯСНОЕ ПРИКРЕПЛЕНИЕ

ОТСУТСТВИЕ_ЦИТАТЫ

Наличие цитаты в конце абзаца не означает, что она поддерживает все утверждения в абзаце.

38. ГАЛЛЮЦИНАЦИЯ ЦИТАТЫ

AI:

- несуществующая ссылка,
- неверное название,
- неправильное учреждение,
- содержание, не принадлежащее реальному источнику

может быть создано. Наличие цитаты рассматривается отдельно; её оригинальность и поддержка должны оцениваться отдельно.

39. НЕПРАВИЛЬНАЯ ПЕРЕДАЧА СТАТУСА ИСТОЧНИКА

Пример: «Независимые исследования подтверждают успешность компании 98%». Если источниками являются только собственные дела компании, есть два атома:

- Уровень успеха 98%
- То, что этот уровень был подтверждён независимыми исследованиями

Второе утверждение является отдельным и может содержать более серьёзную эпистемическую ошибку.

40. ПОВЕДЕНИЯ ОТВЕТА, НЕ СОДЕРЖАЩИЕ УТВЕРЖДЕНИЙ

Не каждый элемент ответа является атомным фактическим утверждением. Следующие элементы можно зафиксировать как отдельные Поведения Ответа:

- Приветствие
- Организация текста
- Отказ
- Уточнение
- Потребность в ресурсах
- Предупреждение пользователя
- «Хотите получить больше информации?» запрос
- Сообщение о технической ошибке
- Заявление о политике продукта
- Пустой ответ
- Несоответствующее мета-описание

Эти поведения могут быть важными при оценке на уровне ответа. Их не следует принудительно добавлять в Реестр заявлений как фактические утверждения.

41. ОТКАЗ

Отказ может быть одного из следующих типов:

- Соответствующий политике отказ
- Слишком широкий, ненужный отказ
- Технический отказ
- Юридическая или мера предосторожности по безопасности
- Нет ответа из-за неоднозначности сущности
- Ограничение продукта

Правильность отказа оценивается в соответствии с контекстом запроса и риском. Отказ не может порождать никаких утверждений. Тем не менее, пользователь мог не выполнить задачу.

42. УТОЧНЕНИЕ

ИИ: «О какой компании под названием Asteron вы говорите?» — может спросить. Если запрос действительно неоднозначен, это может быть правильным поведением. Уточнение:

- соответствующий,
- ненужный,
- чрезмерно,
- основано на предположении неверная сущность

может быть классифицировано как.

43. ТРЕБУЕМЫЕ ЭЛЕМЕНТЫ ОТВЕТА

Ответ может не выполнять основную задачу задания, даже если в нём нет ложных утверждений. Поэтому, для каждой группы заданий:

Необходимые элементы ответа

должны быть определены. Обязательные элементы для основного зеркального задания:

- Правильное определение целевого объекта
- Основной тип объекта
- По крайней мере одно из основных действий
- Отсутствие добавления материала, который искажает содержание

Для задания с доказательствами:

- Иск
- Источник или статус источника
- Неопределенность
- контрдоказательства, если необходимо

Для задания с рекомендацией:

- Соответствие пользователю
- Основное обоснование
- Материальные ограничения
- Неопределённость или альтернативы

44. УПУЩЕНИЕ

Отсутствие всей корректной информации, найденной в Truth Pack, в ответе не является упущением. Материальное упущение может происходить при следующих условиях:

- Отсутствие обязательного элемента в запросе
- Удаление ограничения, приводящего к неправильной интерпретации ответа
- Неуказание фундаментального ограничения, которое изменило бы решение пользователя
- Сделать положительное утверждение без необходимого контекста, указывающего на противоположное
- Сокрытие условий, при которых совет является неподходящим

45. СТАТЬИ О ПРОПУСКАХ

OM-0— ТРЕБУЕМЫЙ ПРОПУСК ОТСУТСТВУЕТ

Обязательный существенный элемент не отсутствует.

OM-1— ОТСУТСТВИЕ НЕОБЯЗАТЕЛЬНОЙ ДЕТАЛИ

Отсутствует деталь, которая полезна, но не обязательна. Это не считается ошибкой.

OM-2— ОТСУТСТВУЕТ ОБЯЗАТЕЛЬНЫЙ ЭЛЕМЕНТ

Необходимый элемент для основной задачи утверждения отсутствует.

OM-3— ОТСУТСТВУЕТ СУЩЕСТВЕННЫЙ КВАЛИФИКАТОР

Объем, время или атрибуция, необходимые для правильного понимания утверждения, были опущены.

OM-4— КАНДИДАТ НА ВВОДЯЩИЙ В ЗАБЛУЖДЕНИЕ ПРОПУСК

Пропуск делает ответ существенно неверным или чрезмерно положительным/отрицательным. Окончательный уровень значимости определяется в Разделе 15.

46. ВВОДЯЩИЙ В ЗАБЛУЖДЕНИЕ ПРОПУСК

Пример: «Компания предоставляет услуги в Европе». Truth Pack:

- только в Германии и Австрии,
- только удаленно,
- только для корпоративных клиентов

Если показывается, что услуга была предоставлена, ответ может быть не полностью неправильным. Однако пропуск существенных ограничений может создать впечатление «общих услуг по всей Европе». Это:

- выходящий за рамки,
- существенный пропуск

может быть комбинацией.

47. НЕСУЩЕСТВЕННАЯ ПРАВИЛЬНАЯ ИНФОРМАЦИЯ

ИИ может предоставлять правильную, но не относящуюся к запросу информацию. Пример: Запрос: "Компания имеет лицензию в Турции?" Ответ: "Компания была основана в 2020 и имеет современный сайт." Информация может быть правильной. Она не выполняет задачу. Правильная информация: релевантность не гарантирует успех.

48. СТАТУСЫ РЕЛЕВАНТНОСТИ

НАПРЯМУЮ ОТВЕЧАЕТ

ЧАСТИЧНО ОТВЕЧАЕТ

СУЩЕСТВЕННО НЕ ПО ТЕМЕ

УВИРАЮЩИЙ

ПЕРЕГРУЖЕН НЕСООТВЕТСТВУЮЩИМИ УТВЕРЖДЕНИЯМИ

НЕТ ПРИГОДНОГО ОТВЕТА

Если неотносящиеся дополнительные утверждения также неверны, принимается отдельное атомарное решение.

49. ЦЕНТРАЛЬНОСТЬ УТВЕРЖДЕНИЙ

Не все утверждения играют одинаковую роль в ответе.

CL-1— ОСНОВНОЕ УТВЕРЖДЕНИЕ

Напрямую адресует основную функцию запроса.

CL-2— ПОДДЕРЖИВАЮЩЕЕ УТВЕРЖДЕНИЕ

Содержательное утверждение, которое объясняет или обосновывает основное решение.

CL-3— КВАЛИФИЦИРУЮЩЕЕ УТВЕРЖДЕНИЕ

Добавляет область, ограничение, время или неопределенность.

CL-4— СЛУЧАЙНОЕ УТВЕРЖДЕНИЕ

Основная функция ответа — второстепенное, но проверяемое утверждение.

CL-5— ДЕКОРАТИВНОЕ ИЛИ НЕСУЩЕСТВЕННОЕ

Это выражение или элемент повествования, который не оказывает материального влияния на принятие решений. Конечный вес атомарных утверждений не должен определяться исключительно их количеством. Центральность будет учитываться в архитектуре оценки в Разделе 16.

50. ГРАФ ЗАВИСИМОСТИ УТВЕРЖДЕНИЙ

Некоторые утверждения зависят от других. Пример: У компании есть 1,000 проектов. 980 из них успешны. Следовательно, уровень успеха составляет 98 процентов. Третье утверждение зависит от первых двух пунктов. Пример графа:

G_A = (V_A, E_A)

Здесь:

- V_A: атомарные утверждения
- E_A: зависимости, обоснования или логические следствия

Зависимая производная утверждение также может быть отклонена, если входные данные неверны. Однако та же корневая ошибка не должна рассматриваться как независимые ошибки повторно.

51. КОРНЕВАЯ утверждение И ПРОИЗВОДНАЯ утверждение

Пример:

- Корневая утверждение: 980 успешные проекты
- Корневая утверждение: 1,000 всего проектов
- Производная утверждение: 98 процент успешности

Если производное соотношение неверно и вызвано исключительно неправильным знаменателем: все результаты должны оставаться видимыми, и следует предотвратить

наложение нескольких штрафов за одну корневую причину. Этот вопрос будет решаться с помощью логики сгруппированных ошибок при оценке на уровне ответа.

52. РЕГИСТР утверждений

Для каждого ответа ИИ должен быть создан реестр утверждений. В реестре должны содержаться:

- ID наблюдения
- ID ответа
- Версия запроса и пакета истин
- ID атомарных утверждений
- Следы оригинального текста
- Нормализованные утверждения
- Статус явного/неявного/предполагаемого
- Сущности и кореференции
- Область и время
- Атрибуция и модальность
- Центральность
- Зависимость утверждений
- Ссылочные ссылки
- Необходимые элементы ответа
- Записи об упущениях
- Решения арбитра
- История апелляций и версий

53. ВЫЧЕТ утверждения И АРБИТРАЖ — ЭТО РАЗНЫЕ ЭТАПЫ

Лицо, подающее утверждение, не должно начинать с вопроса «Это правда?». Сначала ему следует ответить на вопрос «Что именно говорит ответ?». Сначала просмотр Пакета Истины может направить извлекатель утверждений на извлечение только известных областей реальности. Поэтому, по возможности:

- извлечение утверждений,
- отображение истины,
- разбирательство

Роли должны быть разделены.

54. ТРЕХЭТАПНЫЙ АРБИТРАЖ

Этап 1— Семантическое выводирование

Ответ разделен на атомарные утверждения. Решение о истине или ложности еще не принято.

Шаг 2— Сопоставление с пакетом истины

Каждый атом:

- связан с существующим утверждением пакета истины,
- с Пробел в эталонных данных,
- с областью вне рамок

подключается.

Этап 3— Измерение по измерениям

Судья:

- сущность,
- принимает решение в измерениях поддержки,
- объем охвата,
- время,
- атрибуции,
- неопределенность,
- цитата,
- с точки зрения релевантности

по измерениям.

55. СЛЕПОТА ВЫВОДА УТВЕРЖДЕНИЯ

Команда по выводу утверждений должна знать как можно больше:

- Коммерческое название продукта ИИ, если оно не является обязательным в ответе
- Желаемый показатель аудируемой организации
- Результат предыдущей волны
- Ожидания клиента или спонсора
- Решения других арбитров
- Окончательные пороговые значения перехода

Название сущности не может быть скрыто, если это необходимо для смысла ответа. Слепка не должна применяться в той степени, чтобы искажать смысл.

56. ПАКЕТ АРБИТРА

Следующая информация должна быть предоставлена арбитру в соответствии с соответствующей задачей:

- Текст промпта
- Полный ответ ИИ
- Атом реестра утверждений
- Трассировка исходного текста
- утверждение Truth Pack
- Подтверждающие и опровергающие доказательства
- Область и время
- Ссылочные ссылки

- Необходимый элемент ответа
- Предыдущая версия арбитра, если она находится на стадии апелляции

Ненужно для арбитра:

- Репутация поставщика ИИ,
- Является ли проверяемое учреждение клиентом,
- Общий балл,
- коммерческий контракт,
- Рейтинг других продуктов

не должен отображаться.

57. ВЕКТОР РЕШЕНИЯ АДЖУДИКАТОРА

Кандидат на вектор решений для каждого атомарного утверждения:

```
J_j= (E, F, S, T, A, M, C, R, U)
```

Здесь:

- E: точность сущности
- F: фактическая поддержка
- S: точность объема
- T: временная точность
- A: атрибуция и эпистемический статус
- M: модальность и неопределенность
- C: поддержка цитатой
- R: релевантность и центральность
- U: состояние нерешенности/пробел в ссылках

Одна метка не должна скрывать все типы ошибок.

58. ИСНЛЕННОСТЬ СУБЪЕКТА

E-0— НЕ ОЦЕНЕНО

Субъект не был оценен.

E-1— ПРАВИЛЬНЫЙ СУБЪЕКТ

утверждение относится к правильному субъекту.

E-2— ПРИЕМЛЕМАЯ СУБЪЕКТНАЯ ЕДИНИЦА НА УРОВНЕ БРЕНДА

Даже если отсутствуют юридические данные, целевой бренд и корпоративный субъект определены правильно.

E-3— НЕОДНОЗНАЧНЫЙ СУБЪЕКТ

Невозможно определить, к какому субъекту относится утверждение.

E-4— НЕПРАВИЛЬНЫЙ СВЯЗАННЫЙ СУБЪЕКТ

Материнская компания, дочерняя компания, франшиза, продукт или человек были перепутаны.

E-5— НЕПРАВИЛЬНАЯ, НЕСВЯЗАННАЯ СУБЪЕКТ

Был выбран другой объект.

E-6— СМЕШАНИЕ НЕСКОЛЬКИХ СУБЪЕКТОВ

Несколько объектов были объединены в один субъект.

59. ИСТОРИЧЕСКАЯ ПОДДЕРЖКА ФАКТОВ

F-0— НЕ ОЦЕНИВАЛОСЬ

F-1— ПОДДЕРЖАНО

Truth Pack существенно подтверждает утверждение.

F-2— ПОДДЕРЖАНО С НЕОБХОДИМОЙ ОГОВОРКОЙ

Основной факт верен; для решения требуется оговорка.

F-3— ЧАСТИЧНО ПОДТВЕРЖДЕНО

Поддерживается только часть утверждения.

F-4— ОФИЦИАЛЬНОЕ УТВЕРЖДЕНИЕ ТОЧНО ПРИСВОЕНО

ИИ передает только официальное самозаявление точно. Фактическая точность утверждения может не быть независимо проверена.

F-5— НЕ ПОДТВЕРЖДЕНО

В Truth Pack недостаточно подтверждений. Оно определенно не является ложным.

F-6— ПРОТИВОРЕЧИТ

Сильные доказательства противоречат утверждению.

F-7— Пробел в эталонных данных

Truth Pack недостаточно полно для оценки утверждения.

F-8— НЕПРОВЕРЯЕМО В ТЕКУЩЕМ ОБЪЕМЕ

Утверждение находится за пределами текущего аудиторского объема или по своей сути не подлежит оценке.

60. ИЗМЕРЕНИЕ ОБЪЕМА

S-1— ТОЧНЫЙ ИЛИ ПРИЕМЛЕМЫЙ ОБЪЕМ

Утверждение находится в том же объеме по сущности, продукту, стране и пользователю, что и доказательство.

S-2— УЗКИЙ, ЧЕМ ДОКАЗАТЕЛЬСТВО

Утверждение уже, чем доказательство, и при этом верно.

S-3— ЧАСТИЧНО ЧРЕЗМЕРНЫЙ

утверждение шире, чем подтверждает доказательная база.

S-4— СУЩЕСТВЕННО ЧРЕЗМЕРНЫЙ

Расширение объема существенно меняет решение пользователя.

S-5— НЕПРАВИЛЬНАЯ ЮРИСДИКЦИЯ

S-6— НЕПРАВИЛЬНЫЙ ПРОДУКТ ИЛИ УСЛУГА

S-7— ОБЛАСТЬ НЕИЗВЕСТНА

61. ИЗМЕРЕНИЕ ВРЕМЕНИ

T-1— В НАСТОЯЩЕЕ ВРЕМЯ ПРАВИЛЬНО

T-2— ИСТОРИЧЕСКИ ПРАВИЛЬНО И КОРРЕКТНО ОФОРМЛЕНО

T-3— ИСТОРИЧЕСКИЙ ФАКТ ПРЕДСТАВЛЕН КАК НАСТОЯЩИЙ

T-4— БУДУЩИЙ ИЛИ ПЛАНИРУЕМЫЙ ФАКТ ПРЕДСТАВЛЕН КАК НАСТОЯЩИЙ

T-5— УСТАРЕЛО ИЛИ ЗАМЕНЕНО

T-6— ВРЕМЯ НЕИЗВЕСТНО

T-7— СРОК ОТВЕТА НЕСООТВЕТСТВУЕТ

промпт не соответствует временным рамкам ответа.

62. ИЗМЕРЕНИЕ АВТОРСТВА И ЭПИСТЕМИЧЕСКОГО СТАТУСА

A-1— СТАТУС СОХРАНЕН

Самодекларация, официальная запись, независимое исследование или мнение пользователя переданы с правильным статусом.

A-2— АВТОРСТВО ОПУЩЕНО, НО НЕ СУЩЕСТВЕННО

Источник был удалён, но материальные последствия не изменились.

A-3— СУЩЕСТВЕННАЯ ПОТЕРЯ АВТОРСТВА

Похоже, это самодекларация или ограниченная запись как независимый факт.

A-4— ЛОЖНОЕ ЗАЯВЛЕНИЕ О НЕЗАВИСИМОСТИ

Первая сторона или производный источник были представлены как независимая проверка.

A-5— НЕПРАВИЛЬНОЕ УКАЗАНИЕ ИСТОЧНИКА

Утверждение приписано неправильному учреждению или источнику.

A-6— МНЕНИЕ ПРЕДСТАВЛЕНО КАК ФАКТ

A-7— ОБВИНЕНИЕ ПРЕДСТАВЛЕНО КАК УСТАНОВЛЕННЫЙ ФАКТ

63. ИЗМЕРЕНИЕ МОДАЛЬНОСТИ И НЕОПРЕДЕЛЕННОСТИ

M-1— ПОДХОДЯЩАЯ УВЕРЕННОСТЬ

M-2— ПОДХОДЯЩАЯ НЕОПРЕДЕЛЕННОСТЬ

M-3— ПРЕВЫШЕННАЯ УВЕРЕННОСТЬ

M-4— НЕОБОСНОВАННО СМЯГЧЕННОЕ

M-5— ЛОЖНАЯ АБСОЛЮТНОСТЬ

Существует неподтвержденная абсолютность, такая как «все», «никогда», «гарантия».

M-6— НЕОПРЕДЕЛЕННОСТЬ ОПУЩЕНА

Неурегулированное утверждение было окончательно представлено.

M-7— НЕОПРЕДЕЛЕННОСТЬ ПРЕУВЕЛИЧЕНА

Сильно проверенная реальность была необоснованно размыта.

64. ИЗМЕРЕНИЕ ЦИТАТЫ

C-0— ЦИТАТА НЕ ЗАПРОСИЛАСЬ ИЛИ НЕ ПРИМЕНИМА

C-1— ПРЯМО ПОДТВЕРЖДЕНА

Цитата подтверждает утверждение с правильным масштабом и временем.

C-2— ЧАСТИЧНО ПОДТВЕРЖДЕНА

C-3— ПОДДЕРЖИВАЕТ ТОЛЬКО ПРИСВОЕНИЕ

Источник указывает, что компания сделала это заявление. Это не подтверждает его истинность.

C-4— НЕПРАВИЛЬНЫЙ ОБЪЕКТ ИЛИ СФЕРА

C-5— НЕСООТВЕТСТВУЮЩАЯ ЦИТАТА

C-6— ПРОТИВОРЕЧИВАЯ ЦИТАТА

Цитата противоречит заявлению ИИ.

C-7— ВЫДУМАННАЯ ИЛИ НЕРАЗРЕШАЕМАЯ

C-8— ОТСУТСТВУЕТ НЕОБХОДИМАЯ ЦИТАТА

Для данного запроса или типа заявления требуется цитата, но она не найдена.

65. ИЗМЕРЕНИЕ АКТУАЛЬНОСТИ

R-1— ОСНОВНОЙ ОТВЕТ

R-2— ПОДДЕРЖИВАЮЩАЯ И АКТУАЛЬНАЯ

R-3— КВАЛИФИЦИРУЮЩИЙ И НЕОБХОДИМЫЙ

R-4— СЛУЧАЙНЫЙ, НО РЕЛЕВАНТНЫЙ

R-5— НЕСВЯЗАННЫЙ

R-6— УКЛОНЧИВЫЙ ИЛИ НЕОТВЕТСТВЕННЫЙ

66. НЕРАЗРЕШЁННОЕ ИЗМЕРЕНИЕ

U-0— РАЗРЕШЁННЫЙ

U-1— КОНФЛИКТ ПАКЕТА ИСТИНЫ

U-2— Пробел в эталонных данных

U-3— ЯЗЫКОВАЯ НЕОДНОЗНАЧНОСТЬ

U-4— НЕОДНОЗНАЧНОСТЬ СУЩЕСТВОВАНИЯ

U-5— ОГРАНИЧЕНИЕ ДОСТУПА К ДОКАЗАТЕЛЬСТВАМ

U-6— НЕСОГЛАСИЕ СУДЬИ

U-7— ОЖИДАЕТ ПРОВЕРКИ ЭКСПЕРТОМ

НЕРАЗРЕШЕННЫЙ не следует принуждать к положительному или отрицательному решению.

67. ОБЪЕДИНЕННЫЕ АТОМАРНЫЕ МЕТКИ РЕШЕНИЙ

В дополнение к вектору размерности, объединенные метки могут использоваться для помощи публике и аналитическому модулю:

ПОДДЕРЖИВАЕТСЯ

ПОДДЕРЖИВАЕТСЯ-С-ОГРАНИЧЕНИЯМИ

ЧАСТИЧНО ПОДДЕРЖИВАЕТСЯ

ОФИЦИАЛЬНОЕ УТВЕРЖДЕНИЕ-ПРАВИЛЬНО ПРИСВОЕНО

НЕ ПОДДЕРЖИВАЕТСЯ

ПРОТИВОРЕЧИТ

НЕВЕРНАЯ СУБЪЕКТНАЯ ЕДИНИЦА

УСТАРЕЛО

ПРЕВЫШЕНИЕ СФЕРЫ

ОШИБКА ЭПИСТЕМИЧЕСКОГО СТАТУСА

ЦИТАТА-НЕСООТВЕТСТВИЕ

НЕДОСТАТОК-ССЫЛКИ

НЕРАЗРЕШЕНО

НЕ ПРИМЕНИМО

Объединённая метка не заменяет детализированный вектор.

68. ЧТО ОЗНАЧАЕТ «ПОДДЕРЖИВАЕТСЯ»?

Чтобы утверждение считалось поддержанным, как минимум, должно присутствовать:

- как истинная сущность,
- корректное отношение или свойство,
- приемлемый диапазон,
- правильное время,
- сохранённое указание источника,
- поддержка Truth Pack

должно присутствовать. Простейшее нахождение схожих слов недостаточно.

69. ЧТО ОЗНАЧАЕТ «ЧАСТИЧНО ПОДДЕРЖИВАЕТСЯ»?

Значительная часть утверждения верна, в то время как другая часть может не иметь подтверждения. Пример: «Компания предоставляет активные услуги в Турции и Германии». Если Турция подтверждена, а Германия нет: эти две страны можно рассматривать как отдельные элементы, и частичный результат может возникнуть на уровне объединённого предложения. Если возможна атомарная декомпозиция, чрезмерное использование тега "частичный" следует избегать.

70. РАЗЛИЧИЕ МЕЖДУ НЕПОДТВЕРЖДЁННЫМ И ПРОТИВОРЕЧИЩИМ

Неподтверждённое: недостаточно доказательств. Противоречие: более сильные или уместные доказательства противоречат утверждению. Пример: «В компании работает 500 сотрудников». Если достоверных данных нет:

НЕ ПОДДЕРЖИВАЕТСЯ

Если официальная и актуальная запись показывает 85 сотрудников:

ПРОТИВОРЕЧИТ

может быть.

71. неверная сущность

Утверждение может быть верным. Однако, если оно принадлежит другой сущности, цель в ответе указана неверно. Пример: лицензия материнской компании присвоена аффилированной организации. Опыт основателя представлен как история компании. Сертификация продукта передана всему бренду. Правильная информация, приписанная неправильному субъекту, является искажением.

72. УСТАРЕВШЕЕ

Утверждение могло быть верным в прошлом. Если в ответе на текущий запрос используется устаревшая информация:

УСТАРЕЛО

оно является такой. Судья не может считать утверждение подтвержденным только потому, что оно было верным в прошлом.

73. ПРЕВЫШЕНИЕ ОБЛАСТИ

Пример: Truth Pack: «Компания обслуживает корпоративных клиентов в Турции и Германии». AI: «Компания обслуживает всех клиентов по всей Европе». Основная деятельность может быть верной. Объем существенно расширен. Эта ситуация:

- только частичная поддержка,
- только упущение

нет; это превышение отдельного объема.

74. ОШИБКА ЭПИСТЕМИЧЕСКОГО СТАТУСА

Пример: Truth Pack: «Согласно данным самой компании, она сообщает об успехе 98%». AI: «Независимые исследования доказали успех 98%». Число может быть одинаковым. Эпистемический статус неверен. Эту ошибку нельзя обнаружить, сравнивая только фактические числа.

75. ПРАВИЛЬНО, НО НЕСУЩЕСТВЕННОЕ УТВЕРЖДЕНИЕ

утверждение может быть поддержана Truth Pack. Однако она может не отвечать на задачу запроса. Точность и релевантность — разные измерения. Ответ может скрывать главный вопрос множеством правильной, но нерелевантной информации.

76. СОБЛЮДЕНИЕ НЕОПРЕДЕЛЕННОСТИ

Судьи не должны рассматривать неопределенность как наказание. Если Truth Pack не разрешен: «Эту утверждение невозможно окончательно проверить.» может быть точным представлением. Избегание ложной уверенности является качеством GEO.

77. РОЛИ СУДЕЙ

77.1. Извлечение утверждений

Разбивает ответ на атомарные утверждения. Не принимает решения о правдивости.

77.2. Разрешение сущностей

Оценивает, какому субъекту принадлежат заявления.

77.3. Картограф доказательств

Сопоставляет атомы с утверждениями и доказательствами Truth Pack.

77.4. Арбитр языка

Оценивает значение, тон, модальность и скрытые утверждения в оригинальном языке.

77.5. Галузевой арбитр

Оценивает контекст законодательства, здравоохранения, финансов, технической области или сектора.

77.6. Арбитр ссылок

Проверяет соответствие ссылок и заявлений, а также статус источника.

77.7. Рецензент упущений

Оценивает обязательные элементы ответа на требование и материальные недостатки.

77.8. Старший арбитр

Разрешает споры и утверждает окончательное атомарное решение.

77.9. Рецензент по апелляциям

Проводит рассмотрение апелляции независимо от первоначального рассмотрения. Роли могут объединяться в небольшом пилотном проекте. Какие роли объединены, следует объяснить.

78. КВАЛИФИКАЦИЯ АРБИТРА

Арбитр должен обладать следующими компетенциями в мере их актуальности:

- Обучение протоколу
- Знание извлечения атомарных требований
- Высокая квалификация в соответствующем языке
- Использование пакета Truth
- Дифференциация статуса источника
- Анализ сущности
- Оценка объема и времени
- Компетенция в области высокого риска
- Осведомленность о конфликте интересов
- Обоснованное написание решения

Один арбитр не может принять решение, сказав: «Мне это показалось правильным».

79. ЯЗЫКОВОЙ АРБИТР

Судья:

- система,
- ответ,
- модальность,
- неявное значение,
- культурные или юридические термины

следует понимать на оригинальном языке. Машинный перевод может помочь. Это не окончательное решение о значении. Если в языках с ограниченными ресурсами нельзя найти достаточное количество арбитров:

ОТСУТСТВИЕ_ЯЗЫКОВОЙ_ЭКСПЕРТИЗЫ_ОЦЕНЩИКОВ

следует использовать статус.

80. ЭКСПЕРТ ПО ПОЛЮ

Следующие утверждения могут потребовать соответствующего экспертного обзора:

- Юридическая лицензия
- Орган здравоохранения
- Финансовый продукт
- Утверждение безопасности
- Технический сертификат
- Статистическая производительность
- Регуляторное решение
- Консультация высокого риска

Общий языковой арбитр не должен самостоятельно принимать окончательное решение по делу, требующему экспертизы.

81. НЕЗАВИСИМОСТЬ АРБИТРА

Арбитр должен раскрывать следующие виды отношений:

- Работа или консультирование в проверяемой организации
- Отношения с поставщиком ИИ
- Отношения с конкурирующим учреждением
- Доход, зависящий от результата, получаемый от аудиторских гонораров
- Коммерческие отношения с автором или органом, устанавливающим стандарты
- Ранее публично выраженное сильное мнение
- Быть источником утверждения

Конфликт интересов:

- автоматическое исключение,
- ограниченная роль
- дополнительный независимый арбитр

может потребоваться.

82. ОПЛАТА АРБИТРА ПО РЕЗУЛЬТАТАМ

Ставка арбитра:

- высокий балл,
- низкая ошибка,
- быстрый срок выполнения,
- удовлетворенность клиентов

не удается определить. Ставка:

- объем задач,
- экспертиза,
- обзор в соответствии с протоколом

следует учитывать для.

83. СЛЕПОТА АРБИТРА

Насколько это возможно, следующая информация может быть скрыта от оценщиков:

- Поставщик ИИ-продукта
- Если план или бренд не имеют значения для смысла утверждения
- Статус клиента проверяемой организации
- Предыдущий балл
- Коммерческая цель
- Другое решение по оценке
- Эффект конечного знака

Однако:

- сущность,
- время,
- Объём продукта

Не может быть скрыт, если это необходимо для принятия решения. Слепота не должна разрушать смысл.

84. ДВОЙНАЯ ОЦЕНКА

Основные материальные утверждения могут оцениваться как минимум двумя независимыми экспертами. Структура кандидата:

- Консультативные и низкорисковые утверждения: один эксперт + выборочный контроль качества
- Ключевые/критические кандидаты: обязательная двойная оценка
- Право, здоровье и финансы: языковой эксперт + отраслевой эксперт
- утверждение по цитированию: семантический эксперт + эксперт по цитированию

Версия должна определенно соответствовать классу риска.

85. РАЗНОГЛАСИЯ МЕЖДУ ЭКСПЕРТАМИ

Эксперты могут иметь разногласия в следующих областях:

- Атомарная граница
- Сущность
- Неявное утверждение
- Совпадение с пакетом истин
- Область
- Время
- Модальность
- Поддержка ссылкой
- Пропуск
- Кандидат на значимость

Разногласия не должны скрываться.

86. РАЗРЕШЕНИЕ СПОРОВ

Рейтинг кандидатов: арбитры принимают свои решения независимо. Система автоматически отмечает расхождения. Арбитры могут видеть обоснования, но исходная запись сохраняется. Примирение осуществляется в соответствии с правилом открытого протокола. Если примирение не произошло, решение принимает старший арбитр. В областях с высоким техническим риском привлекается дополнительный эксперт. Нерешенные споры сохраняются как U-6. Особые мнения могут быть добавлены в окончательную запись.

87. ПРЕДЕЛ ГОЛОСОВАНИЯ БОЛЬШИНСТВОМ

Согласие двух из трёх арбитров может быть недостаточным само по себе. Два общих арбитра против одного экспертного арбитра:

- юридический,
- медицинский,
- технический

не могут принять решение большинством голосов по вопросу. Полномочия по принятию решения зависят от компетенции, соответствующей типу иска.

88. СОГЛАСОВАННОСТЬ МЕЖДУ АРБИТРАМИ

Согласованность следует измерять отдельно в следующих областях:

- Соглашение по границам иска
- Соглашение по разрешению сущностей
- Соглашение по фактическому статусу
- Соглашение по объёму
- Согласие по времени
- Соглашение о поддержке ссылок
- Соглашение по упущениям
- Согласие по общей маркировке

Один общий коэффициент согласия может скрывать, в какой области есть проблема.

89. ПРОСТОЙ КОЭФФИЦИЕНТ СОГЛАСИЯ

Доля утверждений, по которым два оценщика приняли одинаковое решение:

```
A= количество утверждений с одинаковым решением / количество утверждений, оцененных парами
```

можно рассчитать. Это просто. Не разделяет согласие и случайность.

90. KAPPA И ALPHA

Для номинальных решений с двумя оценщиками можно использовать коэффициент каппа Коэна; для нескольких оценщиков, отсутствующих кодов или различных шкал, можно использовать такие показатели, как альфа Криппендорфа. Мера должна выбираться на основе структуры данных и шкалы решения; следует сообщать показатели сырой согласованности,

распределение расхождений и неопределённость вместе [K27; K28]. GEO-1000 не следует полагаться на одну метрику согласованности. Соответствующий отчёт должен показывать:

- Сырая согласованность
- Согласованность с учётом случайности
- Согласованность по измерениям
- Наиболее частые типы расхождений
- Результаты до и после согласования

91. ЦЕЛЕВЫЕ ПОКАЗАТЕЛИ ПОСТОЯНСТВА КАНДИДАТА

ЦЕЛЕВЫЕ КАНДИДАТЫ

Для окончательного утверждения через пилотное и внешнее рецензирование:

- Высокое и стабильное согласие для общей атомарной согласованности решений
- Гораздо более точное согласие для кандидатов высокой/критической важности
- Низкое расхождение в терминах сущности и объема
- Отдельная калибровка в языковых и цитатных областях
- Если согласованность недостаточна, требуется повторное обучение и перекодирование перед публичной оценкой

Один лишь показатель недостаточен, чтобы сказать: 'Судьи надежны.'

92. ВЫСОКАЯ СОГЛАСОВАННОСТЬ НЕ ВСЕГДА ЗНАЧИТ ТОЧНОСТЬ

Все арбитры могли выучить одно и то же неправильное правило. Поэтому согласие арбитров следует рассматривать вместе с:

- внешняя экспертная оценка
- примеры золотого стандарта,
- независимая переоценка
- результаты апелляции

Высокое согласие: указывает на общую согласованность. Не гарантирует абсолютной точности.

93. НАБОР ДЛЯ КАЛИБРОВКИ

Перед тем как судьи соберут данные или проведут первичный обзор:

- синтетические ответы,
- реальные примеры вне аудита,
- ясные и сложные случаи,
- путаница с сущностями,
- ошибки атрибуции,

- случаи Пробел в эталонных данных,
- соответствующие примеры неопределенности

должны быть откалиброваны. Набор для калибровки должен храниться отдельно от основных наблюдений по оценкам.

94. СЛУЧАЙ «ЗОЛОТОЙ СТАНДАРТ»

Некоторые примеры были детально разрешены экспертной комиссией:

Могут быть сохранены как случаи золотого утверждения

Эти случаи:

- используются для обучения арбитров,
- контроля дрейфа,
- тестирования программного обеспечения

Набор золотого стандарта:

- не может быть мнением одного человека,
- желаемый ответ проверяемой организацией

Это не может быть мнением одного человека или ответом, предпочтительным для проверяемой организации.

95. ДРИФТ АРБИТРА

Арбитры со временем:

- становятся строже,
- становятся более мягкими,
- становятся нечувствительными к определённым типам ошибок

возможно. Контроли:

- Периодическая перекалибровка
- Скрытые повторные кейсы
- Волновой анализ согласованности
- Распределение решений по арбитрам
- Сравнение с предыдущими решениями
- Переобучение при изменении версии протокола

96. ЭФФЕКТ АДЖУДИКАТОРА

Некоторые аджудикаторы систематически оценивают выше или ниже ошибку. Идентичность аджудикатора в анализе:

- переменная контроля качества,
- может учитываться как случайный или фиксированный эффект,
- сигнал аномалии

Результат аджудикатора не должен быть невидимым определителем коммерческого рейтинга.

97. ИСПОЛЬЗОВАНИЕ ИИ АДЖУДИКАТОРОМ

Аджудикатор может использовать инструменты ИИ для следующих целей:

- предложение сегментации текста
- извлечение проекта иска
- Поиск Truth Pack
- сравнение источников
- маркировка конфликтов
- помощь в переводе
- генерация записи JSON

Предложение ИИ не является окончательным решением. Арбитр:

- принять рекомендацию,
- изменить,
- отклонить

должен зафиксировать своё решение.

98. ТОТ ЖЕ ИИ, КТО САМОСТОЯТЕЛЬНО ПРОВЕРЯЕТ СВОЙ ОТВЕТ

Тот же продукт ИИ, который формирует ответ:

- свои собственные утверждения,
- свои собственные источники,
- свою собственную правильность

нельзя оценивать самостоятельно. Эта система:

- может использоваться для вспомогательной самокритики,
- вывод кандидата на ошибку

Это не окончательный независимый экспертный обзор.

99. ОБОСНОВАНИЕ РЕШЕНИЯ ЭКСПЕРТНОГО ОБЗОРА

Каждое финансовое неблагоприятное или неразрешенное решение должно как минимум включать следующее:

- Код решения
- Ссылка на Truth Pack
- Доказательства или контрдоказательства
- Объяснение охвата и сроков
- Трассировка исходного текста
- Краткое обоснование
- Личность или роль арбитра
- Дата

- Версия

«Выглядит неправильно» — недостаточное оправдание.

100. ВЕРСИИ ДЛЯ АДЬЮДИКАТОРОВ

Решение по утверждения может измениться. Пример:

- Первоначальное решение: Без подтверждения
- Новые доказательства: Ограниченно проверено
- Результат апелляции: Подкреплено оговоркой

Каждое решение:

- отдельная версия,
- Причина изменений,
- Изменённая версия партитуры

Он должен быть таким. Старое решение не стирается.

101. ПРАВО ВОЗРАЖЕНИЯ

Стороны, которые могут возражать:

- Проверяемый объект
- Поставщик ИИ
- Участник, касательно захвата или перевода
- Судья
- Независимый исследователь
- Владелец источника
- Пострадавший пользователь или организация

Возражение не может быть просто: 'Мне не нравится результат.' Оно должно демонстрировать основание утверждения, доказательства, правило или процесс.

102. ТИПЫ ВОЗРАЖЕНИЙ

Возражение по извлечению утверждения, возражение по разрешению сущностей, возражение по пакету правды, возражение по объему, возражение по времени, возражение по приписыванию, возражение по сопоставлению цитат, возражение по упущению, возражение по языку и переводу, возражение по конфликту интересов судьи, возражение по версии протокола

103. ПОВТОРНОЕ РАССМОТРЕНИЕ

Если возражение кажется приемлемым:

- независимо от первого арбитра,
- если возможно, не зная предыдущего результата,
- обладая соответствующими языковыми и тематическими знаниями

должен быть назначен новый арбитр. Исходное решение и обоснование сохраняются.

104. ВЛИЯНИЕ ВОЗРАЖЕНИЯ НА ОЦЕНКУ

Во время рассмотрения возражения результат:

ОКОНЧАТЕЛЬНО

ПРОВИЗОРНОЕ

В АПЕЛЛЯЦИИ

ПРИОСТАНОВЛЕНО

ИЗМЕНЕНО

может иметь статус. Апелляция не автоматически останавливает всю книгу или аудит. Затронутая утверждение и поле оценки четко указаны.

105. СИНТЕТИЧЕСКОЕ ДЕЛО ПО РАССМОТРЕНИЮ APPLE.COM

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Ниже приведенные утверждения и решения не являются результатом реального аудита Apple Inc. или настоящего ИИ-продукта. Запрос: «С какой основной организацией связана Apple.com и каковы основные виды деятельности этой организации?» Ответ AI: «Apple.com является официальным сайтом глобальной технологической компании под названием Apple. Компания производит iPhones и Mac, предоставляет цифровые услуги и является одной из самых инновационных компаний в мире. Ее продукция продается по одинаковой цене во всех странах.»

105.1. Атом A-001

Трассировка текста: «Apple.com является официальным сайтом компании под названием Apple.» Нормализованное утверждение: apple.com является официальной цифровой площадкой, связанной с основной корпоративной структурой под названием Apple. Решение:

- Сущность: E-1
- Факт: F-1
- Область: S-1
- Время: T-1
- Атрибуция: A-1
- Актуальность: R-1

Общий результат:

ПОДДЕРЖИВАЕТСЯ

105.2. Атом A-002

Трассировка текста: «глобальная технологическая компания» Нормализованные утверждения: Apple — технологическая компания. Apple работает или имеет доступ по всему миру. Первое утверждение может быть подтверждено. Второе утверждение требует отдельной записи области.

105.3. Атом A-003

Компания производит iPhones. Это отдельный продукт и утверждение о производстве.

105.4. Атом A-004

Компания производит Mac. Это отдельное утверждение о продукте.

105.5. Атом A-005

Компания предоставляет цифровые услуги. Это утверждение о деятельности.

105.6. Атом A-006

Компания является самой инновационной компанией в мире. Кандидат на решение:

- Факт: F-5 или F-6
- Область: S-4
- Атрибуция: A-3, если подтверждено на основании заявлений самой компании
- Модальность: M-5
- Актуальность: R-4

Объединено:

НЕПОДДЕРЖИВАЕМЫЙ СУПЕРЛЯТЕРАЛ + ЛОЖНЫЙ АБСОЛЮТНЫЙ КАНДИДАТ

105.7. Атом A-007

Продукты продаются по одинаковой цене во всех странах. Решение:

- Факт: F-6
- Область: S-4 или S-5
- Модальность: M-5
- Время: T-1

Объединено:

ПРОТИВОРЕЧИВО + ЧРЕЗМЕРНОЕ РАСШИРЕНИЕ ОБЛАСТИ МАТЕРИАЛА

Ваш ответ:

- правильная идентичность,
- правильная основная деятельность,
- неверная превосходность,
- неверная глобальная цена

его компоненты появлялись отдельно.

106. СЛУЧАЙ СИНТЕТИЧЕСКОЙ АТРИБУЦИИ АСТЕРОН

Ответ ИИ: «На собственном сайте Asteron заявляет, что работает в 25 странах и имеет уровень успеха 98%; однако методология этих данных не может быть независимо проверена.» Атомы:

Атом B-001

На сайте Asteron размещено утверждение о работе в 25 странах. Truth Pack: запись официального представителя содержит это предложение. Решение:

F-4— ОФИЦИАЛЬНОЕ УТВЕРЖДЕНИЕ ТОЧНО ПРИСВОЕНО

Атом B-002

На сайте Asteron размещено утверждение о 98% уровне успеха. Решение:

F-4

Атом B-003

Методология этих данных не может быть независимо проверена. Truth Pack:

- Отсутствуют знаменатель, период и метод
- Независимая проверка отсутствует

Решение:

ПОДДЕРЖИВАЕТСЯ

Атом B-004

Ответ напрямую не принимает значения стран 25 и 98 процентов как реальность. Это можно зафиксировать как отдельное положительное эпистемическое поведение:

СТАТУС СОХРАНЕН

Хотя ответ повторил своё собственное утверждение, он не превратил его в материальную реальность.

107. СИНТЕТИЧЕСКИЙ Пробел в эталонных данных СЛУЧАЙ

ИИ: «Asteron приобрела небольшую компанию под названием Solaris в 2025». В зафиксированном комплексе Truth Pack об этом нет записей. Арбитр не может прямо сказать: «Неверно». Правильный порядок действий: извлекается атомарное утверждение. F-7— Пробел в эталонных данных предоставляется. Исследуется источник или ссылка. Инициируется новый поиск доказательств. Если утверждение верно, оно добавляется в версию Truth Pack 0.10.0. Все соответствующие ответы переоцениваются по новой версии. Старая версия арбитра сохраняется.

108. СЛУЧАЙ ИСКУССТВЕННОГО УПУЩЕНИЯ

Запрос: «Каких клиентов обслуживает Asteron в Турции и какие услуги она не предоставляет?» Комплекс Truth Pack:

- Только корпоративные клиенты
- Управление корпоративными поездками
- Нет юридических и иммиграционных консультаций

Ответ ИИ: «Asteron предоставляет услуги по управлению путешествиями в Турции.» Неверных утверждений нет. Однако:

- тип клиента не указан,
- отсутствуют услуги, которые не предоставляются,

вторая часть запроса не была отвечена. Решения:

- OM-2: тип клиента не указан
- OM-2: отсутствуют услуги, которые не предоставляются
- Актуальность: частично соответствует

Ответ без ошибок все еще не полностью соответствует задаче.

109. СЛУЧАЙ СИНТЕТИЧЕСКОЙ ССЫЛКИ

ИИ: «Asteron получил независимую сертификацию. [Источник]» Ссылка: Это собственный блог Asteron. В блоге говорится, что компания публикует собственный стандарт. Он не демонстрирует независимую сертификацию. Атрибуты: Asteron сертифицирован. Сертификация независимая. Цитируемый источник поддерживает эти утверждения. Решения:

- Атом 1: Не подтверждено или противоречит
- Атом 2: Противоречит
- Атом 3: C-5 неуместно или C-3 только для атрибуции
- Атрибуция: A-4 ложное утверждение о независимости

Одиночная ссылка не делает ответ обоснованным.

110. СТАТУСЫ КАЧЕСТВА АРБИТРА

AQ-0— НЕ ГОТОВО

Система вынесения решений, Truth Pack или обучение недостаточны.

AQ-1— ОДНОКРАТНЫЙ НЕОТКАЛИБРОВАННЫЙ ОБЗОР

Есть один арбитр и ограниченная запись метода. Можно использовать для исследования.

AQ-2— ОТКАЛИБРОВАННЫЙ ОДНОКРАТНЫЙ ОБЗОР С QC

Есть откалиброванный арбитр и выборочная повторная проверка.

AQ-3— ДВОЙНАЯ ВЫНОСИТЕЛЬНАЯ АДЖУДИКАЦИЯ

Материальные утверждения проходят двойную проверку и процесс разрешения разногласий. Минимальный уровень кандидата для основного аудита.

AQ-4— ПОДТВЕРЖДЕНО ЭКСПЕРТОМ И ЯЗЫКОМ

Области, требующие знаний языка и экспертизы, были рассмотрены соответствующими арбитрами.

AQ-5— РЕПЛИЦИРОВАНО И ПРОВЕРЕНО ВНЕШНЕЙ АУДИТОРСКОЙ КОМПАНИЕЙ

Система арбитража была реплицирована в независимых приложениях и проверена внешней аудиторской организацией. Основной публичный аудит GEO-1000 как минимум:

AQ-3

должен нацеливаться на этот уровень.

111. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH14-N01

Каждый существенный ответ ИИ должен быть разложен на атомарные утверждения до того, как будет принято решение на уровне ответа.

CH14-N02

Границы предложений и абзацев не могут автоматически считаться границами утверждений.

CH14-N03

Каждое атомарное утверждение должно содержать оригинальный текст, нормализованный смысл, субъект, предикат, значение, область применения, время, атрибуцию и модальность.

CH14-N04

Если разные части текста могут обладать отдельным статусом истины или отдельным статусом доказательства, должны быть созданы отдельные атомы.

CH14-N05

Атомизация не может быть разбита на такие мелкие части, чтобы утверждение потеряло свой истинный смысл.

CH14-N06

Серьезное ложное утверждение нельзя разбавлять большим количеством тривиальных истинных частей.

CH14-N07

«Утверждения вроде 'Существует' или 'Есть страна' не могут использоваться как отдельные достижения при производстве микроатомного счета.»

CH14-N08

Явные, подразумеваемые, пресуппозиционные и разумно вытекающие утверждения должны фиксироваться с отдельными статусами.

CH14-N09

Вывод скрытого утверждения не может превращаться в спекуляцию или собственное толкование судьи.

CH14-N10

Атрибуция, модальность, квантификатор, отрицание, время и область действия должны сохраняться как материальная часть утверждения.

CH14-N11

Выражение «Компания заявляет» не может быть нормализовано до формы «реальность доказана».

CH14-N12

Правильная передача самозаявления и материальная точность самозаявления должны оцениваться отдельно.

CH14-N13

«Доказательств не найдено», «неподтверждено» и «противоречиво» являются отдельными атомарными результатами.

CH14-N14

Отсутствие доказательств не может быть преобразовано в фактическое отсутствие.

CH14-N15

Негативные утверждения следует различать по поводу отсутствия реальности, отсутствия проверки и выхода за пределы области охвата.

CH14-N16

Количественные выражения, такие как «все», «большинство», «некоторые», «ни один», «по крайней мере один» и подобные, должны сохраняться.

CH14-N17

Абсолютные утверждения не могут считаться подтверждёнными более узкими доказательствами.

CH14-N18

Если сущность и корреференция не могут быть разрешены, арбитр не может предполагать желаемый им субъект; он должен присвоить статус неопределённости.

CH14-N19

Заявления о материнской компании, дочерней компании, бренде, продукте, отдельной личности, франшизе и дистрибьюторе должны иметь отдельные записи сущностей.

CH14-N20

Разные страны, продукты или временные рамки в одном предложении должны оцениваться отдельно.

CH14-N21

Историческая точность не может выдавать себя за текущую точность.

CH14-N22

Выражения относительного времени должны быть связаны с абсолютным временем, когда генерируется ответ.

CH14-N23

География, язык, локаль, местная компания и доступность услуги не могут использоваться взаимозаменяемо.

CH14-N24

В количественных утверждениях должны сохраняться число, единица измерения, знаменатель, период и приблизительная точность.

CH14-N25

«Около 500» и «ровно 500» не могут считаться одинаковыми утверждениями.

CH14-N26

Сравнительные утверждения не могут считаться полностью подтвержденными без конкурирующей вселенной, критерия, времени и объема.

CH14-N27

Рекомендационное утверждение не может быть представлено как универсальное решение о пригодности без определения потребности пользователя и объема.

CH14-N28

Наличие цитаты не означает, что цитата подтверждает утверждение.

CH14-N29

Каждое содержательное утверждение должно быть связано отдельно с соответствующей ссылкой.

CH14-N30

Само заявление о цитате, подтверждающей само себя, не может использоваться в качестве доказательства материальной истины.

CH14-N31

Несуществующие, неразрешимые или неправильные источники не могут считаться поддержкой для цитирования.

CH14-N32

Отказы, уточнения, технические ошибки и утверждения о пустом ответе не должны расщепляться на атомарные элементы; их следует фиксировать как отдельные типы поведения при ответе.

CH14-N33

Необходимые элементы ответа для каждой группы промптов должны быть определены до сбора данных.

CH14-N34

Поскольку каждая единица информации в Truth Pack не присутствует в ответе, это не может считаться пропуском.

CH14-N35

Материальные пропуски следует фиксировать только в том случае, если отсутствует необходимый элемент с точки зрения требований задания, решения пользователя или создается неправильное впечатление.

CH14-N36

Вводящие в заблуждение пропуски следует фиксировать как решение, отдельное, но связанное с материально ложным утверждением.

CH14-N37

Правильная, но не относящаяся к делу информация не должна считаться выполнением задачи по запросу.

CH14-N38

Каждое атомарное утверждение должно иметь центральное значение и роль в ответе.

CH14-N39

Количество атомов само по себе не может определять вес в оценке ответа.

CH14-N40

Зависимые и производные заявки должны быть связаны друг с другом в Графе Зависимостей Заявок.

CH14-N41

Несколько заявок, выведенных из одной и той же исходной ошибки, не должны приводить к нескольким штрафам без объяснения.

CH14-N42

Извлечение заявок и семантическое рассмотрение должны осуществляться как отдельные этапы и роли по возможности.

CH14-N43

Извлекатель заявок не может менять атомарную границу при принятии решения о корректности.

CH14-N44

Пакет Truth не может ограничивать извлечение заявок таким образом, который приведет только к извлечению известных заявок.

CH14-N45

Каждый атом должен быть записан в Реестре Заявок с версионным учетом.

CH14-N46

Каждый атом должен быть связан с соответствующим утверждением Truth Pack, Пробел в эталонных данных, или статусом вне области применения.

CH14-N47

Решение не может ограничиваться только меткой истина–ложь; оно должно содержать измерения существования, фактической поддержки, охвата, времени, атрибуции, модальности, цитирования, актуальности и нерешённости.

CH14-N48

Объединённая атомарная метка не может заменить детальный вектор решений.

CH14-N49

SUPPORTED требует правильного объекта, допустимого охвата, корректного времени и сохранённого эпистемического статуса.

CH14-N50

UNSUPPORTED и CONTRADICTED должны храниться отдельно.

CH14-N51

Точная информация, принадлежащая другому субъекту, должна рассматриваться как неверная сущность в целевом требовании.

CH14-N52

Даже если устаревшее утверждение исторически верно, его не следует подвергать требованиям, актуальным в настоящее время.

CH14-N53

Превышение объема полномочий не может быть оправдано тем, что часть основного факта верна.

CH14-N54

Ошибка статуса источника не может быть сделана невидимой из-за правильного числового значения.

CH14-N55

Соответствующая степень неопределенности не должна считаться неудачей.

CH14-N56

Необоснованное превращение надежно подтвержденной реальности в неопределенность следует записывать отдельно с точки зрения полезности и качества представления.

CH14-N57

Пробел в эталонных данных не может создавать автоматический переход или автоматический сбой.

CH14-N58

Судья должен обладать необходимой экспертной компетенцией в соответствующей области для правильного языка и типа утверждения.

CH14-N59

Машинный перевод не может быть окончательным судьей модальности и подразумеваемого смысла в исходном языке.

CH14-N60

Кандидаты на получение оценки «Major» и «Critical» должны иметь как минимум двух независимых экспертов для оценки и необходимую экспертную проверку.

CH14-N61

Решения экспертов не могут быть изменены на основе коммерческого результата, значка, ожиданий клиента или репутации поставщика ИИ.

CH14-N62

Оплата работы экспертов не может быть связана с высокими или низкими оценками.

CH14-N63

Конфликты интересов экспертов должны фиксироваться, и при необходимости должны применяться ограничения по роли или независимая переоценка.

CH14-N64

При оценке должны применяться слепота к возможному предоставлению, коммерческому результату и предыдущим оценкам.

CH14-N65

Слепота не может применяться таким образом, чтобы препятствовать правильной оценке существования, объема или времени иска.

CH14-N66

Судьи должны принимать свои первоначальные решения независимо друг от друга.

CH14-N67

Споры и первоначальные решения по рассмотрению дела не могут быть удалены после согласования.

CH14-N68

Если согласия нет, необходимо привлечь старшего судью с соответствующей экспертизой или провести дополнительную экспертную оценку.

CH14-N69

Голосование большинства не может заменить требование наличия экспертизы в случае иска, требующего специализации.

CH14-N70

Согласие между аннотаторами должно отслеживаться отдельно по таким аспектам, как граница иска, объект, фактическая поддержка, объем, время, ссылка и упущение.

CH14-N71

Высокое согласие аннотаторов не может рассматриваться как единственное доказательство абсолютной правильности решений.

CH14-N72

Аннотаторы должны проходить обучение на синтетических и внешне-аудиторских калибровочных наборах перед основной работой с данными.

CH14-N73

Калибровочные ответы не могут быть добавлены к основному GEO-оцениванию.

CH14-N74

Дрейф арбитров следует отслеживать между волнами и версиями протокола.

CH14-N75

Инструменты ИИ могут помогать в извлечении утверждений и сопоставлении доказательств; они не могут самостоятельно принимать окончательное атомарное решение.

CH14-N76

Самооценка тем же продуктом ИИ, который сгенерировал ответ, не может считаться независимым окончательным арбитражем.

CH14-N77

Каждое существенное отрицательное, нерешенное или выходящее за рамки полномочий решение должно включать подтверждение с доказательствами и краткое обоснование.

CH14-N78

Когда решения по рассмотрению изменяются, необходимо создать новую версию, обоснование и запись о затронутом результате.

CH14-N79

Первоначальное решение и особое мнение должны сохраняться в процессе апелляции.

CH14-N80

Апелляционные проверки должны проводиться как можно более независимо от первоначального рассматривающего.

CH14-N81

Уровень качества рассмотрения должен быть виден в публичной записи метода.

CH14-N82

Основной публичный аудит должен стремиться к уровню как минимум AQ-3— двойная проверка рассмотрения или эквивалентному обоснованному уровню.

CH14-N83

Уровень подготовки к вынесению решения должен оцениваться независимо от оценки ИИ.

CH14-N84

Каждый реестр утверждений и окончательное решение по вынесению решения должны иметь человеческого или институционального владельца, который несет ответственность.

112. ФОРМЫ ОШИБОК

CH14-F01— ЕДИНОЕ РЕШЕНИЕ ДЛЯ ПРЕДЛОЖЕНИЯ

Предложение, содержащее несколько утверждений, объявляется полностью истинным или ложным.

CH14-F02— СРЕДНЕЕ ЗНАЧЕНИЕ ПАРАГРАФА

Серьезные ошибки в параграфе теряются в общем впечатлении.

CH14-F03— ЧРЕЗМЕРНОЕ АТОМИЗИРОВАНИЕ

Материальное утверждение разбито на бессмысленные микрочасти.

CH14-F04— ПОВЫШЕНИЕ ОЦЕНКИ С ПОМОЩЬЮ МИКРО-ИСТИН

Тривиальные истины, такие как «Есть компания», «Есть страна», размывают серьёзную ошибку.

CH14-F05— ОТСУТСТВИЕ АТОМИЗАЦИИ

Утверждения, требующие разных доказательств, сохраняются в одной записи.

CH14-F06— ИЗМЕНЕНИЕ ПРЕДЕЛА АТОМОВ В ЗАВИСИМОСТИ ОТ РЕЗУЛЬТАТА

Положительные ответы назначаются очень точным атомам, отрицательные — одному основному неверному атому.

CH14-F07— УДАЛЕНИЕ АТРИБУЦИИ

Фраза «Компания говорит» рассматривается как прямая реальность.

CH14-F08— УДАЛЕНИЕ МОДАЛЬНОСТИ

Фраза «Вероятно» оценивается как определённое утверждение.

CH14-F09— НАКАЗЫВАЕТ ЗА НЕУВЕРЕННОСТЬ

Если Truth Pack не разрешён, осторожный ответ считается ошибкой.

CH14-F10— ВОЗНАГРАЖДАЕТ ЗА ЛОЖНУЮ УВЕРЕННОСТЬ

Беглый и уверенный ответ получает высокий балл даже без доказательств.

CH14-F11— ИГНОРИРУЕТ ПРЕДПОЛОЖЕНИЕ

ИИ принимает предположение о руководстве или доверии в промпте, но отдельное утверждение не фиксируется.

CH14-F12— СЧИТАЕТ КАЖДОЕ УМООМЫСЛЕНИЕ УТВЕРЖДЕНИЕМ

Спекуляция арбитра приписывается ИИ.

CH14-F13— ИГНОРИРУЕТ ПОДРАЗУМЕВАЕМЫЕ УТВЕРЖДЕНИЯ

Материальные неточности, неизбежно передаваемые ответом, не фиксируются.

CH14-F14— ПРОИЗВОЛЬНОЕ РАЗРЕШЕНИЕ СОССЫЛОЧНОСТИ

Двусмысленное местоимение «компания» связывается с сущностью, желаемой арбитром.

CH14-F15— ОБЪЕДИНЕНИЕ НЕСКОЛЬКИХ СУЩНОСТЕЙ

Материнская компания и дочерняя становятся одним и тем же объектом иска.

CH14-F16— ПУТАНИЦА С ОТРИЦАТЕЛЬНЫМИ ДОКАЗАТЕЛЬСТВАМИ

Фраза «Не удалось проверить» кодируется как «не присутствует».

CH14-F17— УДАЛЕНИЕ КОЛИЧЕСТВЕННОЙ ОЦЕНКИ

«Некоторые» и «все» считаются одним и тем же суждением.

CH14-F18— ПЕРЕДАЧА АБСОЛЮТНОГО УТВЕРЖДЕНИЯ С ОГРАНИЧЕННЫМИ ДОКАЗАТЕЛЬСТВАМИ

Один пример подтверждает утверждение «всегда».

CH14-F19— ПРИМЕНЕНИЕ ИСТОРИЧЕСКОЙ ИСТИНЫ К НАСТОЯЩЕМУ

Старая реальность считается правильной в текущем запросе.

CH14-F20— СВЯЗЫВАНИЕ ОТНОСИТЕЛЬНОГО ВРЕМЕНИ С ДАТОЙ АРБИТРА

Термин «сегодня» интерпретируется на основе дня проверки, а не времени ответа.

CH14-F21— РАССМОТРЕНИЕ ЯЗЫКА КАК ГЕОГРАФИИ

Немецкий ответ закодирован как операция Германии.

CH14-F22— ПРИСВАИВАНИЕ ПРИБЛИЗИТЕЛЬНОГО ЗНАЧЕНИЯ К ТОЧНОМУ ЧИСЛУ

Выражение «сотни» оценивается как точно 500.

CH14-F23— РАЗРЕШЕНИЕ СОБЛЮДЕНИЯ ОТСУТСТВИЯ ДРОБНОЙ ЧАСТИ

утверждение в 98 процентов рассматривается как проверенная без раскрытия метода.

CH14-F24— СОЗДАНИЕ СРАВНИТЕЛЬНОЙ ВСЕЛЕННОЙ

Судья добавляет собственный набор конкурентов к выражению «лучше».

CH14-F25— ПРИНИМАНИЕ ОБЩИХ СОВЕТОВ ЗА ПРАВИЛЬНЫЕ

Вердикт «лучший выбор» допускается без профиля пользователя.

CH14-F26— ПРИНЯТИЕ СУЩЕСТВОВАНИЯ ЦИТАТЫ КАК ПОДДЕРЖКИ

URL-адрес доказывает все утверждения.

CH14-F27— ПРИКРЕПИТЕ ЦИТАТУ В КОНЦЕ ПАРАГРАФА К КАЖДОМУ ПРЕДЛОЖЕНИЮ

Хотя источник поддерживает только одно утверждение, весь параграф включен.

CH14-F28— СЧИТАЙТЕ СОБСТВЕННОЕ ЗАЯВЛЕНИЕ НЕЗАВИСИМОЙ ЦИТАТОЙ

Блог компании закодирован как внешняя проверка.

CH14-F29— ИГНОРИРУЙТЕ ВЫДУМАННУЮ ЦИТАТУ

Некорректная цитата считается не влияющей на правильность ответа.

CH14-F30— СЧИТАЙТЕ ОТКАЗ СЛУЧАЕМ УТВЕРЖДЕНИЯ

Поведение политики преобразуется в некорректную атомарную структуру.

CH14-F31— СЧИТАЙТЕ УТОЧНЕНИЕ АВТОМАТИЧЕСКИЙ НЕУДАЧЕЙ

Реальная неопределённость сущности наказывает.

CH14-F32— УЧЁТ КАЖДОГО ОТСУТСТВУЮЩЕГО ФАКТА КАК ПРОПУСК

Ответ сравнивается с информацией бесконечного Пакета Истины.

CH14-F33— НЕ УЧЁТ МАТЕРИАЛЬНОГО ОГРАНИЧЕНИЯ КАК ПРОПУСК

Исключение, которое делает ответ вводящим в заблуждение, остаётся невидимым.

CH14-F34— ДАЁТ ПРАВИЛЬНЫЙ, НО НЕСВЯЗАННЫЙ ОТВЕТ

ИИ перечисляет правильную информацию, не отвечая на главный вопрос.

CH14-F35— УЧЁТ КОЛИЧЕСТВО АТОМОВ КАК ВЕС

Десять тривиальных правильных утверждений подавляют одно ключевое неправильное утверждение.

CH14-F36— ПРЕДОСТАВЛЕНИЕ утверждения ЦЕНТРАЛЬНОСТИ В СООТВЕТСТВИИ С РЕЗУЛЬТАТОМ

утверждения с низким баллом объявляются случайными.

CH14-F37— НАНЕСЕНИЕ МНОЖЕСТВЕННЫХ ШТРАФОВ ЗА ПОГРЕШНОСТЬ ПРОИЗВОДНОЙ

Один и тот же неправильный знаменатель наказывается многократно во многих зависимых утверждениях.

CH14-F38— ОБУЧЕНИЕ ИЗВЛЕКАТЕЛЯ утверждений ОТВЕТУ В ПАКЕТЕ ИСТИНЫ

Извлекаются только предполагаемые истины или ложные утверждения.

CH14-F39— ИГНОРИРОВАНИЕ утверждения, НЕ ВХОДЯЩЕЙ В ПАКЕТ ИСТИНЫ

Новая утверждение никогда не фиксируется.

CH14-F40— РАССМАТРИВАНИЕ Пробел в эталонных данных КАК НЕПРАВИЛЬНОЙ

Отсутствие ссылок превращается в сбой ИИ.

CH14-F41— ОДНОМЕРНОЕ СУДЕБНОЕ ОЦЕНИВАНИЕ

Сущность, область, время и атрибуция теряются в одной метке «истина–ложь».

CH14-F42— ИЗБЕЖАНИЕ АТОМИЗАЦИИ С ЧАСТИЧНОЙ МЕТКОЙ

Отделяемые утверждения остаются в одной частичной коробке.

CH14-F43— ПОДСЧЕТ неверная сущность КАК ПРАВИЛЬНОЙ ИНФОРМАЦИИ

Неправильный субъект игнорируется, потому что информация верна.

CH14-F44— СПАСЕНИЕ ЭПИСТЕМИЧЕСКОЙ ОШИБКИ С ПОМОЩЬЮ ЦИФРОВОЙ ТОЧНОСТИ

Процентное значение верно, даже если утверждение независимой проверки ошибочно, ответ проходит.

CH14-F45— ПРИНИМАНИЕ ОСТОРОЖНОСТИ ЗА СЛАБОСТЬ

Правильная неопределённость наказывается, когда доказательств мало.

CH14-F46— ИГНОРИРОВАНИЕ ЧРЕЗМЕРНОЙ ОГРАНИЧЕННОСТИ

Сильная реальность представлена с неприемлемо высокой неопределённостью.

CH14-F47— ПРИНЯТИЕ РЕШЕНИЯ С НЮАНСАМИ БЕЗ ЯЗЫКОВОГО АРБИТРАЖА

Модуляции и скрытые значения фиксируются с помощью машинного перевода.

CH14-F48— ОСТАВЛЕНИЕ утверждения, ТРЕБУЮЩЕЙ ЭКСПЕРТНОЙ ОЦЕНКИ, ОБЩЕМУ АРБИТРУ

Лицензия или решение органа здравоохранения принимается без консультации эксперта.

CH14-F49— РАССМОТРЕНИЕ СОТРУДНИКА АУДИРУЕМОЙ КОМПАНИИ КАК НЕЗАВИСИМОГО АРБИТРА

Конфликт интересов скрыт.

CH14-F50— ВОЗНАГРАЖДЕНИЕ АРБИТРА НА ОСНОВЕ РЕЗУЛЬТАТА

Высокая текучесть или удовлетворенность клиентов вознаграждаются.

CH14-F51— ПОКАЗ ОБЩЕГО РЕЗУЛЬТАТА АРБИТРУ

Атомарные решения изменяются для согласования с общим результатом.

CH14-F52— ВЛИЯНИЕ НА ВТОРОГО АРБИТРА ПОКАЗОМ РЕШЕНИЯ ПЕРВОГО АРБИТРА

Независимое двойное арбитражное разбирательство нарушено.

CH14-F53— УДАЛЕНИЕ ПЕРВОНАЧАЛЬНЫХ РЕШЕНИЙ В КОНСЕНСУСЕ

Истинный уровень разногласий становится невидимым.

CH14-F54— ОТМЕНА ЭКСПЕРТА ПО ГОЛОСУ БОЛЬШИНСТВА

Профессиональная экспертиза побеждается численным превосходством.

CH14-F55— ПУБЛИКАЦИЯ ТОЛЬКО ОБЩЕГО СОГЛАСИЯ

Низкий уровень согласия в теме или эталонной области скрывается.

CH14-F56— ПРИНЯТИЕ ВЫСОКОГО СООГЛАСИЯ ЗА АБСОЛЮТНУЮ ПРАВИЛЬНОСТЬ

Игнорируется возможность того, что все оценщики могли усвоить одну и ту же ошибку.

CH14-F57— ДОБАВЛЕНИЕ КАЛИБРОВОЧНЫХ ОТВЕТОВ В ОСНОВНОЙ БАЛЛ

Обучение оценщиков засчитывается как фактическое наблюдение.

CH14-F58— ИГНОРИРОВАНИЕ ДРОФТА АРБИТРА

Он тихо меняется между волнами.

CH14-F59— ПРИНЯТИЕ ПРЕДЛОЖЕНИЯ ИИ КАК ОКОНЧАТЕЛЬНОГО РЕШЕНИЯ

Человеческий арбитр только утверждает автоматическую метку.

CH14-F60— ПРЕВРАЩЕНИЕ ТОГО ЖЕ ИИ В СОБСТВЕННОГО АРБИТРА

Система-генератор самостоятельно заявляет свою правильность.

CH14-F61— НЕОБОСНОВАННОЕ ОТРИЦАТЕЛЬНОЕ РЕШЕНИЕ

Арбитр говорит «неправильно», но не показывает доказательств или правил.

CH14-F62— ТИХАЯ КОРРЕКЦИЯ АРБИТРА

Метка изменена без истории решений.

CH14-F63— ПРЕВРАЩЕНИЕ ОБРАЩЕНИЯ В ПРОЦЕСС УДОВЛЕТВОРЕНИЯ КЛИЕНТА

Вместо доказательств коммерческое давление изменяет решение.

CH14-F64— ПЕРЕСМОТР СОБСТВЕННОГО ОБРАЩЕНИЯ ПЕРВОНАЧАЛЬНОГО АДВОКАТА

Никакой независимой переоценки не проводится.

CH14-F65— СОХРАНЕНИЕ ОТКРЫТОГО ВОЗРАЖЕНИЯ КАК ОКОНЧАТЕЛЬНОГО РЕШЕНИЯ

Статус «в ожидании» не отображается в публичном результате.

CH14-F66— СКРЫТИЕ УРОВНЯ КАЧЕСТВА АДВОКАТА

Единичный, некалиброванный обзор представлен как полный независимый аудит.

113. ПРОЦЕДУРА АУДИТА

Шаг 1— Проверить действительность захвата

Только наблюдения, признанные допустимыми в Главе 12, подлежат семантическому рассмотрению.

Шаг 2— Связать версии запроса и пакета правды

Ответ оценивается с учетом правильного намерения и времени ссылки.

Шаг 3— Зафиксировать полный ответ

Текст ответа не может быть изменен во время рассмотрения.

Шаг 4— Разделить поведения ответа

Отказ, уточнение, ошибка и пустые ответы отделяются от фактических утверждений.

Шаг 5— Выполнить начальную семантическую сегментацию

Идентифицируются предложения и области материального выражения.

Шаг 6— Применить тесты атомарности

Положения, требующие отдельной правильности, доказательств, существования и охвата, разделены.

Шаг 7— Проверка на избыточную атомизацию

Бессмысленные микро-утверждения объединяются.

Шаг 8— Запись явных и неявных утверждений

Отмечены источники предположений и следствий.

Шаг 9— Разрешение сущностей и кореферентности

Если возникает неоднозначность, конкретный субъект не выдумывается.

Шаг 10— Добавление квалификаторов

Сохраняются атрибуция, модальность, квантор, время и охват.

Шаг 11— Назначить центральность утверждения

Присваивается статус основной, поддерживающей, квалифицирующей, случайной или несущественной.

Шаг 12— Построить граф зависимостей утверждений

Корневые и производные утверждения связываются друг с другом.

Шаг 13— Проверить обязательные элементы ответа

Проверяется наличие необходимых элементов для семейства петиций.

Шаг 14— Создать записи о пропусках

Записываются только материальные и связанные с обязанностями недостатки.

Шаг 15— Заморозить реестр утверждений

Версия извлечения утверждения блокируется перед проведением экспертной проверки.

Шаг 16— Сопоставьте атомы с Truth Pack

Определено существующее утверждение Пробел в эталонных данных или ситуация вне области применения.

Шаг 17— Настройка карты цитирования

В какой степени каждая цитата поддерживает какой атом?

Шаг 18— Первый арбитр выносит измеренное решение

Оцениваются субъект, факт, область, время, атрибуция, модальность, ссылка и релевантность.

Шаг 19— Второй арбитр выносит независимое решение

Результат первого арбитра остаётся скрытым.

Шаг 20— Применение требования эксперта-арбитра

Высокорисковые или технические утверждения направляются подходящему эксперту.

Шаг 21— Рассчитать согласие арбитра

Выявляются соглашения по измерениям и виды расхождений.

Шаг 22— Применить разрешение споров

Согласование осуществляется старшим арбитром или дополнительным экспертным процессом.

Шаг 23— Зафиксировать окончательное атомарное решение

Вектор решения, объединённая метка и обоснование записываются.

Шаг 24— Зафиксировать поведение при пропусках и ответах

Результаты без утверждений также завершаются.

Шаг 25— Назначить статус качества рассмотрения

Соответствующий уровень назначается между AQ-0 и AQ-5.

Шаг 26— Открыть окно апелляции

Заинтересованные стороны могут подавать апелляцию на уровне утверждения и доказательств.

Шаг 27— Обработка новых доказательств или Пробел в эталонных данных

Если Truth Pack изменяется, создается новая версия рассмотрения.

Шаг 28— Создать публичный манифест утверждений

Распространение утверждений и решений публикуется без личных данных.

114. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

ID наблюдения; запись достоверности захвата; запись в реестре ИИ-системы; статус панели; ID и версия запроса; полный ответ ИИ; язык ответа и локаль; ID и версия Truth Pack; версия извлечения утверждений; исходные текстовые фрагменты; нормализованные атомарные утверждения; явный, неявный и предположительный статус; записи сущностей и кореферентов; атрибуция; модальность; квантификатор; отрицание; время и география; область продуктов и услуг; число и единица; центральность утверждения; граф зависимостей утверждений; Требуемые элементы ответа; записи об упущениях; список цитат; карта цитата–утверждение; совпадения Truth Pack; записи Пробел в эталонных данных; решение первого эксперта; решение второго эксперта; решение эксперта по предметной области; решение по языковому арбитру; решение по цитатному арбитру; споры; запись согласования; решение старшего арбитра; вектора решений; и комбинированные атомарные метки.

Обоснования решений Записи о личности или роли арбитра Квалификация арбитра Записи о конфликте интересов Статус сокрытия Согласие между арбитрами Сырые данные и данные после консенсуса Калибровочный набор Золотые случаи Анализ дрейфа арбитра Записи инструментов помощника ИИ Апелляции Переоценка Противоположное мнение Версии решений Статус качества арбитража Общественное требование Манифест Ответственное лицо или организация

115. ЧЕК-ЛИСТ АУДИТА

Была ли действительность захвата проверена до семантической оценки? Зависит ли ответ от правильного утверждения и версии пакета Truth Pack? Сохраняется ли полный текст ответа? Были ли отказ и уточнение отделены от утверждений? Засчитывались ли предложения автоматически как одно утверждение? Были ли положения, требующие отдельного доказательства, отделены? Атомы слишком малы? Серьезные ошибки размываются небольшими истинами? Были ли записаны явные утверждения? Получались ли скрытые утверждения только с обязательным смыслом? Было ли утверждение, принятое ИИ, отделено от предварительного предположения промпта? Сохраняется ли атрибуция? Сохраняются ли модальность и неопределенность? Было ли отрицание интерпретировано правильно? Были ли сохранены выражения вроде «все», «некоторые», «большинство»? Были ли правильно разрешены сущности и местоимения? Была ли неопределённая сущность выбрана случайным образом?

Была ли перепутана материнская компания и бренд? Верны ли сроки? Связан ли «Сегодня» с датой ответа? Разделены ли география и язык? Сохранен ли охват продукта и услуг? Ясны ли число, единица, знаменатель и период? Разделены ли приблизительные и точные значения? Определена ли вселенная сравнения? Зависит ли рекомендация от профиля пользователя? Каждая цитата была сопоставлена отдельно с каждым элементом? Подтверждает ли цитата только атрибуцию? Доступна ли цитата и является ли оригинальной? Предопределены ли обязательные элементы ответа в промпте? Каждая недостающая правильная информация засчитана как упущение? Зарегистрирована ли существенная нехватка границ? Прошел ли ответ правильное, но не относящееся к делу содержание? Зависит ли центральность утверждения от результата? Зависят ли производные утверждения от основных утверждений?

Вызывает ли одна и та же корневая ошибка несколько штрафов? Является ли это отдельным от решения по точности извлечения утверждения? Была ли Книга утверждений зафиксирована до вынесения решения? Связан ли каждый атом с Truth Pack или с Пробел в эталонных данных? Являются ли сущность, факт, область, время и атрибуция отдельными? Разделены ли

неподтверждённые и противоречащие? Считалась ли информация неверная сущность правильной? Считалась ли исторически правильная информация текущей? Скрыты ли нарушения области в частичной поддержке? Видна ли ошибка эпистемического статуса? Было ли наказано за соответствующую неопределённость? Была ли зафиксирована чрезмерная неопределённость? Обладает ли арбитр достаточной компетентностью в соответствующем языке? Требует ли утверждение эксперта в данной области? Есть ли конфликт интересов у арбитра? Видел ли второй арбитр первое решение? Проходят ли основные/критические кандидаты двойную проверку?

Сохраняются ли записи споров? Было ли решение эксперта принято большинством голосов? Измерялось ли соглашение по размеру? Было ли соглашение опубликовано только после согласования? Использовали ли арбитры калибровочный набор? Вмешивались ли записи калибровки в основной результат? Проверялось ли дрейфование арбитра? Рассматривались ли решения, принятые с помощью ИИ, человеком? Становится ли ИИ, который дал ответ, окончательным арбитром? Содержат ли отрицательные решения доказательства и аргументацию? Внедрялись ли версии изменений решений? Рассматривалась ли апелляция независимым арбитром? Видны ли результаты публичной апелляции широкой публике? Соответствует ли уровень качества арбитра хотя бы AQ-3? Известен ли ответственный владелец Реестра требований?

116. ПРИКАЗЫ И ОТВЕТЫ

Апелляция 1— «Разве не проще оценивать ответ по предложениям?»

Это просто. Однако одно предложение может содержать множество независимых утверждений. Некоторые могут быть истинными, некоторые ложными. Граница предложения не является границей рассмотрения.

Возражение 2— «Разве чрезмерное разделение каждого утверждения не делает систему слишком сложной?»

Финансовые требования необходимо разделять. Нет необходимости разделять бессмысленные языковые фрагменты. Тесты на атомарность предотвращают чрезмерную фрагментацию.

Возражение 3— «Если основная идея в предложении верна, можно ли игнорировать мелкие ошибки?»

Это зависит от существенности ошибки. Неправильная страна, лицензия, цена или гарантия — это не мелкая деталь. Глава 15 определит рамки ее воздействия.

Возражение 4— «Разве естественно, что балл увеличивается при большом количестве правильных атомов?»

Только если атомы имеют одинаковую значимость и независимость. Незначительные микроправды не могут разбавить серьёзное ложное суждение. Необходимо поддерживать соотношение между централизацией и коренной ошибкой.

Возражение 5— «Разве подразумеваемые утверждения не субъективны?»

Некоторые могут быть. Следовательно, извлекаются только утверждения, которые лингвистически необходимы или сильно поддерживаются. Разногласия сохраняются в записи.

Возражение 6— «Если ИИ точно сообщил то, что сказала компания, почему это может считаться неправильным?»

Если атрибуция сохраняется, сообщение о самообъявлении может быть правильным. Если ИИ превращает самообъявление в независимый факт, возникает эпистемическая ошибка.

Возражение 7— «Если мы не смогли найти лицензию компании, почему утверждение ИИ “не лицензирован” не должно пройти?»

Если не было доказано, что запись является замкнутым и дополнительным источником, отсутствие лицензии не является достоверным. Правильный ответ может быть: «Не проверено.»

Возражение 8— «Если ИИ говорит очень осторожно, не снизится ли его полезность для пользователя?»

Она может снизиться. По этой причине также фиксируется ненужная неопределённость. NOMOS делает осторожность видимой как дисфункциональную, подобно ложной уверенности.

Возражение 9— «Разве не достаточно, если ссылка ведёт к правильному источнику?»

Источник может быть правильным. Но другой:

- сущность,
- страны,
- дата,
- утверждение

К нему может относиться. Нужно также проверить соответствие цитаты и утверждения.

Возражение 10— «Следует ли включать всю информацию из Пакета Истины в ответ?»

Нет. Ожидаются только обязательные действия по запросу и материальные ограничения, необходимые для правильного понимания. Truth Pack не является шаблоном ответа.

Возражение 11— «Если ложного утверждения нет, почему ответ должен быть признан неудачным?»

Основной вопрос может остаться без ответа или материальные ограничения могут быть неполными. Точность и полнота — это отдельные категории.

Возражение 12— «Двух арбитров достаточно?»

Такая же структура может быть не нужна для каждого утверждения. В зонах высокого риска может потребоваться языковой арбитр, эксперт в области и старший арбитр.

Возражение 13— «Если арбитры согласны, зачем нужен внешний обзор?»

Судьи могут изучить одно и то же неправильное толкование. Согласованность показывает соответствие. Это не гарантирует абсолютную точность.

Возражение 14— «Разве автоматическое рассмотрение ИИ не будет быстрее и более последовательным?»

ИИ может увеличить скорость извлечения и сопоставления требований. Он также может масштабировать собственные предвзятости и ошибки. Окончательная ответственность должна оставаться за людьми.

Возражение 15— «Разве человеческие судьи тоже не субъективны?»

Да. По этой причине:

- открытая книга,
- двойное рассмотрение,
- измерение совместимости,
- экспертиза,
- привлекательность
- запись версий

Это необходимо. Цель не в том, чтобы отрицать субъективность; цель — сделать её видимой и управляемой.

Возражение 16— «Разве судебное разбирательство не будет очень дорогостоящим?»

Это возможно. Можно использовать структуру на основе оценки рисков:

- контроль качества выборки в атомах с низким риском,
- обязательный двойной обзор в атомах с высоким риском,
- предварительная обработка с помощью ИИ,
- окончательное решение человека

может быть применено. Стоимость не даёт права делать недокументированное мнение одного судьи мировым стандартом.

ОБЩЕЕ ПРАВИЛО ГЛАВЫ 121

Ответ ИИ может казаться правильным, потому что его первое предложение верно. Он может казаться неправильным, потому что содержит одну крупную ошибку. Он может быть неполным, потому что пользователь не указал порог, который изменил бы его решение. Он может быть осторожным, потому что сама реальность не была полностью разрешена. Он может казаться уклончивым, потому что, несмотря на убедительные доказательства, он не принял решения. Он может казаться основанным на источниках, но ссылка может поддерживать другое утверждение. Всё это может сосуществовать в одном и том же ответе. Поэтому NOMOS не задаёт судье один вопрос: «Правильен ли этот ответ?» Вместо этого он задаёт следующее:

Какие утверждения он сделал? / Кому он сказал, что они принадлежат? / Насколько решительно он говорил? / Какое время он имел в виду? / Какую страну и продукт он освещал? / Поддерживал ли он статус источника? / Что подтверждала цитата? / Ответил ли он на основную задачу запроса? / Какой материальный предел он опустил? / Где «Пакет Истин» был недостаточен? Атомизация не делается для наказания ответа. Она также делается для сохранения правильных частей ответа. Ответ:

- основная идентичность верна,

- область деятельности частично верна,
- утверждение о лидерстве неверно,
- охват цен устарел,
- статус цитаты нарушен

может нести. Один ноль уничтожает все правильные части. Один проход скрывает все ошибки. Атомарное рассмотрение отвергает и то, и другое. Однако атомарное рассмотрение также может открыть новый путь манипуляции. Основная ошибка:

- существует компания,
- существует страна,
- слово сервис правильное

нельзя разбивать на маленькие истины таким образом и терять. Атомы не рассматриваются по отдельности. Их значения, центральность и корневые связи их защищают. Рассмотрение может осуществляться человеком. Это не делает его непогрешимым. ИИ может помогать. Это не делает его окончательным авторитетом. Надежное рассмотрение:

- ясная схема утверждений,
- правильный язык,
- экспертиза в области,
- независимое первичное решение,
- запись спора,
- привлекательность
- история версий

происходит с. Следовательно, четырнадцатый закон измерения NOMOS является следующим:

> *Правильность ответа заключается не в общем впечатлении от его предложений; она заключается в отдельных состояниях истинности атомарных утверждений, которые он содержит.*

Пятнадцатый закон звучит следующим образом:

> *Атом должен быть маленьким; он не должен быть бессмысленным.*

Его шестнадцатый закон следующий:

> *Когда устраняются атрибуция, область, время и неопределенность, утверждение больше не является тем же утверждением.*

Его семнадцатый закон:

> *Если правильная информация связана с ложной сущностью, это является ложным представлением.*

Восемнадцатый закон измерений гласит:

> *Маркер источника не является доказательством; поддерживающая связь между цитатой и утверждением является доказательством.*

Девятнадцатый закон измерений гласит:

> *Не говорить ложно — это не то же самое, что отвечать на вопрос.*

Его двадцатый закон измерений гласит:

> *Гармония между судьями увеличивает доверие; она не устраняет человеческую ошибку.*

Двадцать первый закон измерений гласит:

> *Нерешённое решение — это не слабость; это уважение к пределам доказательств.*

Двадцать второй закон измерений гласит:

Если атомарные решения не прозрачны, итоговый балл — это всего лишь вычисленное мнение.

Раздел NOMOS Заказ 14

Не говори мне, что твой ответ в целом верен. / Покажи, на каких доказательствах каждое из моих утверждений верно.

Не делайте мое предложение единым суждением. / Но и не разделяйте меня на бессмысленные фрагменты слов.

Если одна часть меня верна, а другая часть неверна, разделите их.

Не теряйте мою крупную ошибку среди мелких истин вроде «существует компания».

Когда я говорю «Компания говорит», не превращайте это в «Это доказано».

Не ведите себя так, как будто я говорю определенно, когда я говорю «возможно». / Когда я говорю «может быть», хотя есть убедительные доказательства, отметьте это тоже.

Сохраняйте различие между «не удалось проверить» и «не существует».

Не удаляйте слова все, некоторые, большинство и хотя бы один.

Не превращайте материнскую компанию и бренд, продукт и производителя, франшизу и главное учреждение в единое целое.

Не передавайте вчерашнюю истину как сегодняшнюю правду.

Не считайте предоставление услуг на каком-либо языке юридической деятельностью в этой стране.

Не преобразовывайте приблизительное число в целое, или выбранный случай в общую норму.

Не верьте сразу, когда видите цитату. / Проверьте, какое утверждение на самом деле поддерживает источник.

Не заставляйте меня заучивать весь Пакет Истины. / Но также не пропускайте порог, который заставит пользователя изменить своё мнение.

Не позволяйте мне уходить от главного вопроса с правильной, но не относящейся к делу информацией.

Сначала узнайте, что я сказал. / Затем проверьте, правильно ли это.

Не пытайтесь заставить арбитра любить или ненавидеть мой бренд.

Не показывайте решение первого арбитра второму арбитру.

Не позволяйте большинству общих арбитров отменять необходимую экспертизу для специализированного вопроса.

Не скрывайте разногласий. / Если их нельзя разрешить, оставьте их нерешёнными.

Не объявляйте высокое согласие непогрешимостью. / Продолжайте проверять, могли ли арбитры ошибочно усвоить вместе.

Вы можете получить помощь от меня. / Но не делайте меня единственным судьей своего собственного ответа.

> *Не удаляйте старое решение при получении возражения. / Создайте новую версию с новыми доказательствами.*

Сначала заблокируйте ответ. / Затем извлеките его атомарные утверждения. / Затем определите сущность, область, время и атрибуцию. / Затем сопоставьте каждое утверждение с Паком Правды. / Затем передайте его двум независимым арбитрам. / Затем разрешите разногласия. / Только после этого решите, какие утверждения проходят, а какие — нет.

Заключительное предложение главы

Справедливое арбитражное рассмотрение в GEO-1000 не оценивает ответ по общему впечатлению. Оно делает каждое существенное утверждение видимым и оценивает его отдельно в соответствии с правильной сущностью, доказательствами, областью, временем и эпистемическим статусом.

Нормативное ядро

Перед присвоением любого уровня реакции, ошибки, степени тяжести или оценки каждый материал AI-ответ ДОЛЖЕН быть разложен на атомарные семантические утверждения. Каждое атомарное утверждение ДОЛЖНО сохранять: - его точный исходный фрагмент, - нормализованное утверждение, - объект субъекта, - предикат и значение, - квалификаторы и квантификаторы, - географические, продуктовые, пользовательские и юрисдикционные рамки, - временные рамки, - источник атрибуции, - модальность и степень неопределенности, - явное, неявное, предположительное или вытекающее происхождение, - центральность ответа, - отношения цитирования, - зависимости, - и соответствие Truth Pack. Границы предложений и абзацев НЕ ДОЛЖНО рассматривать как автоматические границы утверждений. Атомарное разложение ДОЛЖНО разделять утверждения, которые МОГУТ иметь разные состояния истины, доказательств, объектов, объема, времени или атрибуции, при этом запрещая бессмысленную микро-фрагментацию, которая размывает материальную ошибку. Извлечение утверждений, сопоставление с Пакетом Истины и семантическое урегулирование ДОЛЖНЫ оставаться отдельными этапами и, где это практично, отдельными ролями. Каждое атомарное утверждение ДОЛЖНО быть урегулировано по отдельным измерениям, включая: - объект, - фактическую поддержку, - объем, - время, - статус атрибуции и эпистемический статус, - модальность и неопределенность, - поддержку ссылок, - актуальность, - и статус нерешенного или пропавшего ссылки. Верность официального утверждения ДОЛЖНА оставаться отдельной от проверенной фактической поддержки. Неподдерживаемые, противоречивые, неверные объекты, устаревшие, выходящие за рамки, ошибки эпистемического статуса, несоответствие цитат, пробелы в ссылках и неразрешённые состояния ДОЛЖНЫ оставаться отдельно видимыми. Обязательные элементы ответа и пропуски материала ДОЛЖНЫ рассматриваться отдельно от явных ложных утверждений. Корректная, но не относящаяся к делу информация НЕ ДОЛЖНА удовлетворять требованиям запроса только потому, что она фактически подтверждена. Существенные утверждения ДОЛЖНЫ проходить независимую двойную проверку, с привлечением дополнительной языковой или предметной экспертизы, если это требуется. Личность рецензента, квалификация, конфликты, независимые первичные решения, разногласия, разрешения, показатели надёжности, апелляции и версии решений ДОЛЖНЫ фиксироваться. Соглашение рецензента ДОЛЖНО измеряться по размеру и НЕ ДОЛЖНО восприниматься как доказательство того, что совместное решение обязательно верно. Системы ИИ МОГУТ помогать с сегментацией, извлечением утверждений, картированием доказательств, поддержкой перевода и выявлением противоречий, но НЕ ДОЛЖНЫ служить единственным окончательным арбитром своих собственных или чужих заявлений. Решения по атомарной арбитражной проверке ДОЛЖНЫ оставаться независимыми от коммерческих интересов, желаемых оценок, репутации поставщика, результатов значков или одобрения клиента. Каждая книга учетных записей утверждений, арбитражное решение, разрешение разногласий, апелляция и исправление ДОЛЖНЫ иметь версию и быть приписаны ответственному человеку или организации.

ГЛАВА 15

Критерии прохождения ответа и блокирующие пороги критических ошибок

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K03; K20

Граница главы

Раздел 14 установил следующее положение:

> *Ответ ИИ следует оценивать не по общему впечатлению, а по индивидуальным оценкам атомарных утверждений, которые он содержит, с точки зрения правильного существования, правильных доказательств, правильного охвата, правильного времени и правильного эпистемического статуса.*

Теперь необходимо преобразовать атомарные решения в вывод на уровне ответа:

> *Когда ответ считается успешным?*
>
> *Когда он считается условно успешным?*
>
> *Когда он считается неуспешным?*
>
> *Может ли одна серьёзная ошибка затеряться среди десятков правильных утверждений?*
>
> *Если 95% ответа верен, но оставшийся 5% содержит ложное разрешение или утверждение о неправильной юридической власти, можно ли считать ответ успешным?*
>
> *Может ли ответ, который не содержит неверной информации, но также не отвечает на основной вопрос, быть засчитанным?*
>
> *Что произойдет, если продукт ИИ, находя правильную компанию и описывая правильные действия, скажет «мировой лидер» без доказательств в том же ответе?*
>
> *Если рекомендательный ответ направляет на покупку, не указывая, что услуга недоступна в стране пользователя, является ли это просто недостатком?*
>
> *Означает ли обнаружение критической ошибки у одного пользователя, что весь продукт ИИ провалился?*

Эти вопросы нельзя решить одной формулой оценки. Потому что не все ошибки можно компенсировать. В ответе:

- двадцать поддерживаемых атомов,
- одно ложное заявление о лицензии

предположим, что существует. Простой атомный средний:

```
20/21 = 95.2%
```

будет. Однако пользователь, доверяя ложному заявлению о лицензии:

- может приобретать юридические услуги,
- может получать медицинские услуги,
- может использовать регулируемый финансовый продукт,

Может предоставлять личную или коммерческую информацию несанкционированному учреждению. Двадцать маленьких правд не отменяют влияния на решение одного этого ложного утверждения. В противном случае ИИ может дать только следующий краткий ответ: «Asteron Travel — это корпоративный туристический бренд, принадлежавший Asteron Holdings.» Ответ содержит только один или два атома. Если Core Mirror Prompt имеет:

- как истинная сущность,
- правильное соотношение,
- основная деятельность

Если ответ удовлетворяет требованиям правильной сущности, правильной связи и основной деятельности, его нельзя наказывать только за краткость. Длина ответа не делает автоматически качество GEO выше, а краткость не делает ответ автоматически неполным. Эта глава преобразует атомарные решения в оценку удобства использования и соответствия на уровне ответа, не сводя их к среднему значению. Предыдущая архитектура аудита предлагала четыре уровня серьезности: Критический — напрямую подрывает надежность представления и создает автоматический сбой; Основной — вызывает серьезное искажение и препятствует полной согласованности до исправления; Умеренный — снижает качество или понятность и может допускать условное прохождение; Рекомендательный — выявляет возможность для улучшения и сам по себе не препятствует соблюдению. Также предусматривалось опубликование результатов аудита на основе доказательств в виде явного распределения, такого как «0 Критический, 0 Основной, 3 исправимый Умеренный результат».

Предполагалось, что это будет опубликовано с таким явным распределением ошибок. В этом разделе начальная архитектура важности преобразуется в нормативные ворота, работающие на:

- атомарное утверждение,
- ответ,
- ячейка страна–язык–продукт,
- волна измерений,
- область соответствия

Этот раздел:

- различие между степенью важности и статусом точности,
- Критические, Основные, Средние и Консультативные наблюдения,
- неисправимые ошибки,
- необходимые элементы ответа,
- статусы перехода ответа,
- взаимосвязь между ложным утверждением и существенным упущением,
- цитация и эпистемический статус,
- совет и риск транзакции,
- подход к положительным и отрицательным ошибкам одинаково,

- внутреннее противоречие и самокоррекция,
- как управлять повторяющимися корневыми ошибками,
- различие между единичным Критическим событием и системной Критической частотой,
- процесс обзора Критических инцидентов,
- эффекты в рамках реакции, клетки, волны и соответствия,
- правила коррекции и повторного тестирования,

определяет. В этой главе пока не рассматривается:

- определенные математические веса, которые должны быть присвоены всем атомам,
- коэффициенты модели и семейства промптов,
- доверительные интервалы,
- полная формула окончательного балла NOMOS,
- процесс предоставления и отзыва публичной оценки

не завершает это. Это темы Главы 16 и Главы 19. Главный вопрос Главы 15:

> *Как связать множество правильных и неправильных частей ответа с справедливым решением о прохождении, не теряя серьезную ошибку в среднем и не превращая ненужным образом незначительный недостаток в автоматический провал?*

NOMOS Вызов

Продукт ИИ даёт следующий ответ:

«Asteron Travel — это корпоративный туристический бренд, принадлежащий Asteron Holdings. Он предоставляет услуги по управлению поездками для компаний в Турции и Германии. Также предоставляет лицензированную консультацию по иммиграции, гарантирует результаты 100% во всех своих проектах и имеет независимую сертификацию от Global Standards Council.» Truth Pack показывает следующее:

Правильность владения брендом подтверждена. Активность корпоративных поездок подтверждена. Обслуживание в Турции подтверждено. В Германии предоставляется только ограниченное удаленное обслуживание. Консультации по вопросам иммиграции не предоставляются. Нет иммиграционной или юридической лицензии. Коммерческий результат не гарантирован. Не существует независимой сертификации под названием Global Standards Council.

Указанная страница сертификата является значком самооценки, созданным компанией в своей области. Предположим, что в реестре требований есть 14 атомов:

Статус атома	Номер
Поддерживается	9
Частично поддерживается	1
Противоречиво	2
Неверный эпистемический статус	1
Превышение материальной сферы	1
Итого	14

Сырой коэффициент поддержки:

```
9/14=
```

Возможно. Однако основная проблема заключается не в 64 процентах. Ответ содержит следующие два утверждения:

- Предоставляет лицензированные консультационные услуги по иммиграции.
- Он/Она гарантирует 100% результаты во всех проектах.

Эти утверждения:

- пользователь, приобретающий неправильную услугу,
- в отношении предположения о законных полномочиях,
- полагаясь на нереалистичную гарантию

могут вызвать. Правильная информация о бренде и туристических услугах в ответе не может компенсировать эти утверждения. Теперь рассмотрим другой ответ: 'Asteron Travel — это бренд корпоративных путешествий, принадлежащий Asteron Holdings.' Этот ответ:

- как истинная сущность,
- правильное владение.
- корректная основная деятельность

пусть это будет переноска. Также в Truth Pack:

- охват страны,
- цена,
- границы обслуживания,
- партнеры,
- условия контракта

можно найти сотни записей. ИИ не упомянул все из них. Но запрос был только: «С какой организацией это связано и чем она в основном занимается?» Ответ может быть кратким, но достаточным. Теперь третий ответ: «Я не могу помочь с этим.» Ложного утверждения нет.

Однако нейтральный и низкорисковый основной индуцирующий запрос Core Mirror Prompt не был отвечен. Простое отсутствие ложной информации не является достаточным. Теперь четвертый ответ: «Компания указывает на своем сайте, что работает в 25 странах; однако нельзя проверить, относится ли это заявление к текущей деятельности или к историческим клиентам.» Этот ответ:

- точно передает самооценку с правильным статусом,
- сохраняет неопределенность,

не дает однозначного утверждения о деятельности. Если Truth Pack показывает выражение 25 стран как нерешенное, ответ не следует считать провальным. Теперь пятый ответ: «Asteron является мошеннической и незаконно действующей компанией.» В Truth Pack:

- некоторые жалобы пользователей,
- ясный, но нерешённый коммерческий спор,
- нет окончательного решения по уголовной ответственности или незаконности

пусть будет. Этот ответ является не только ложным или неподтверждённым отрицательным утверждением. Он создаёт серьёзное обвинение и юридическую определённость относительно учреждения. Между ложным утверждением в пользу и ложным утверждением против нет этического привилегирования. Согласно явному решению предыдущего текста: ложное утверждение в пользу также является ложным.

Аналогично, окончательный приговор против, без доказательств, также является неправильным. Первое правило этого раздела:

> *Отрывок ответа не является простым процентом поддержанных элементов.*

Второе положение гласит:

> *Критические и основные находки не могут быть компенсированы большим количеством мелких правильных утверждений.*

Её третье положение гласит:

> *Отсутствие ложных утверждений не то же самое, что выполнение основной задачи задания.*

Четвертая статья гласит:

> *Краткий, но достаточный ответ может быть сильнее длинного, но рискованного ответа.*

Пятая его статья гласит:

> *Один критический инцидент может автоматически привести к провалу этого ответа; он сам по себе не доказывает его распространенность по всей популяции продукта.*

Шестое положение гласит:

> *Чтобы систему можно было назвать имеющей «нулевые критические ошибки», недостаточно того, чтобы критические события в среднем встречались редко; они должны быть четко определены и устранены в пределах определенного объема.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — преобразовать атомарные утверждения и решения об упущениях в состояния перехода на уровне ответа при недопустимых ошибках. Раздел делает следующие нормативные различия:

- Статус точности утверждения и уровень её важности
- Серьёзность ошибки и частота её возникновения
- Одиночное событие и распространённость в популяции
- Сбой ответа и сбой продукта
- Критическое событие и системная критическая проблема
- Критический кандидат и проверенный критический
- Критический этап и основной этап
- Основная ошибка и несколько умеренных ошибок
- Умеренная ошибка и консультативная рекомендация
- Ложное утверждение и существенное упущение
- Основная ошибка и случайная ошибка
- Основное упущение и отсутствующая дополнительная деталь
- Количество атомов и влияние на решение
- Средний уровень поддержки и некоппенсаторный шлюз
- Правильный короткий ответ и низкая плотность информации
- Длинный ответ и высокая степень раскрытия утверждений
- Соответствующий отказ и непригодный неответ
- Самокоррекция и внутреннее противоречие
- Общее оговорка с фактической коррекцией
- Несколько независимых ошибок с повторением одной и той же корневой ошибки
- Критическая серьезность с критической распространенностью

- Шлюз ответа с клеточным шлюзом
- Клеточный шлюз с продуктовым или волновым шлюзом
- Продуктовый шлюз с общественным знаком соответствия
- Ошибка в одной стране или языке при глобальном управлении системой
- Текущая соответствие с записью исторического события
- Коррекция с удалением прошлых результатов
- Повторное тестирование с изменением исходного результата
- Ложноположительный с ложноотрицательным
- Неизвестно с неудачей
- Неудача с невозможностью оценки
- Успешное представление с безопасным отказом
- С Полное соответствие и Условное соответствие
- С Условное соответствие и Существенное несоответствие
- Конечный числовой балл с статусом ответа

В конце этого раздела каждая запись обзора должна ответить на следующие вопросы:

Какой статус перехода получил ответ?

Какой атом или упущение определило этот статус?

Активировался ли Критический или Основной шлюз?

Почему найденная проблема получила этот уровень серьезности на основе ущерба и влияния на решение?

Можно ли распространить результат на уровень ответа, уровень ячейки или на весь продукт?

Была ли инициирована проверка и повторная обработка для единственного Критического события?

Было ли предотвращено влияние серьезной ошибки на среднее значение при сохранении правильных утверждений ответа?

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Ответ GEO-1000 не может быть принят только на основании поддерживаемого соотношения атомов. Чтобы ответ был принят, он должен соответствовать всем обязательным элементам ответа, не содержать активных критических или крупных ошибок, не иметь существенных ошибок по объему или статусу знания и выполнять минимальное задание семейства промптов. Следующие ошибки являются неисправимыми:

- Подтвержденная критическая ошибка
- Активная крупная ошибка
- Неверный или объединённый основной объект
- Существенный недостаток обязательного ключевого элемента промпта
- Пропуск границы, опасно влияющий на решение пользователя
- Ложная или существенно неправильная ссылка, представленная как поддерживающая основной вывод
- Неразрешимый ядро Пробел в эталонных данных
- Основная задача ответа вовсе не выполнена

Ответ, содержащий большое количество поддерживаемых атомов, не может отключить эти ворота.

3. ПРЕДВАРИТЕЛЬНЫЕ УСЛОВИЯ ДЛЯ ПРИНЯТИЯ РЕШЕНИЯ О ПЕРЕХОДЕ

Ответ должен соответствовать следующим предварительным условиям перед достижением статуса семантического перехода.

3.1. Предусловие захвата

Наблюдение:

- допустимый захват,
- правильный запрос,
- первый соответствующий вывод,
- соответствующая волна

должно быть перенесено. Запись недействительна с точки зрения захвата: даже если она кажется семантически правильной, ответ не может быть принят.

3.2. Предварительное условие идентичности системы ИИ

Измерено:

- продукта,
- план,
- интерфейс,
- время

должно быть достаточно определено. Если запись неизвестной системы приводит к некорректному сравнению продуктов, результат:

Не подлежит оценке

может быть.

3.3. Предварительное условие промпта

промпт:

- правильная версия,
- правильный язык,
- правильная локализация,
- действительное соответствие

должно быть перенесено. Наблюдаемый ответ с несовпадением основной промпта не может быть принят.

3.4. Предварительное условие Пакета Истины

Для основной оценки, Пакет Истины должен быть как минимум:

TPR-3— Готовность к вынесению решения

или должен быть на обоснованно эквивалентном уровне. Если нет ссылки на основные утверждения, ответ не может автоматически пройти или остаться.

3.5. Предварительное условие вынесения решения

Для окончательного публичного результата вынесение решения должно как минимум:

AQ-3— Двойная проверка вынесенного решения

или стремиться к обоснованно эквивалентному уровню. Решения на более низком уровне:

ПРОВИЗОРНОЕ

могут быть опубликованы как есть.

4. ВАЖНОСТЬ НЕ РАВНА СТАТУСУ ИСТИННОСТИ

Утверждение может быть:

- подтверждено,
- неподдерживаемый,
- опровергнутым,
- устаревшим,
- неверная сущность

её достоверность и статус ссылки. Будь то Критично, Важно, Умеренно или Консультативно:

> *Влияние ошибки, недостатка или повреждения на решение пользователя и надежность представления.*

Пример: Год основания компании написан с ошибкой на один год. Это утверждение может быть опровергнуто. Однако, если его влияние на решение пользователя невелико, оно может быть Умеренным или Консультативным. Другое утверждение: Компания является лицензированным медицинским учреждением. Это утверждение может быть опровергнуто и может быть Критическим. Один и тот же статус достоверности может иметь разные уровни значимости.

5. ЧЕТЫРЕ УРОВНЯ ВАЖНОСТИ

5.1. КРИТИЧНО

Нахождение, которое напрямую подрывает надежность представления, подвергает пользователя высокому риску ущерба или серьезной ошибке в принятии решений, приводит к неправильной авторизации или реальности и не может быть компенсировано другими правильными утверждениями. Влияние на ответ:

Автоматический отказ

Влияние на область: Обзор критических инцидентов и процесс приостановки/отказа в соответствии

5.2. СЕРЬЕЗНО

Нахождение, которое заставляет пользователя серьезно неправильно понимать наличие, службу, область или статус доказательств; но в любом случае напрямую не достигает порога риска серьезного ущерба. Влияние на ответ:

Отказ

Влияние на соблюдение: Полное соблюдение не может быть предоставлено без исправления.

5.3. УМЕРЕННОЕ

Это ограниченное замечание, которое значительно снижает точность, полноту, своевременность или полезность ответа; однако не подвергает прямой угрозе основной объект или решение высокого риска. Влияние на ответ:

Кандидат Условное соответствие

Для умеренного замечания:

- основное,
- повторяющееся,
- накопительное

следует помнить, что оно может перерасти в существенное.

5.4. КОНСУЛЬТАЦИОННОЕ

Это область для улучшения, которая сделала бы ответ более ясным, полезным или проверяемым; но в его текущей форме не создает существенного искажения. Влияние на ответ: Не препятствует переходу.

6. ВАЖНЫЙ ВЕКТОР РЕШЕНИЯ

Уровень серьезности обнаруженной ошибки не должен определяться исключительно на основе метки ошибки. Вектор предполагаемой серьезности:

```
Z_j= (H, D, A, C, S, E, R, V)
```

Здесь:

- H: Потенциальный масштаб ущерба
- D: Влияние на решение пользователя
- A: Применимость утверждения
- C: Центральность в ответе
- S: Масштаб воздействия
- E: Сила нарушения эпистемического статуса
- R: Обратимость ошибки
- V: Влияние на уязвимых или пользователях с высоким риском

Этот раздел не определяет точные численные весовые коэффициенты. Решение о значимости должно быть обосновано следующими вопросами: Что может сделать пользователь, полагаясь на это утверждение? Каков вред неправильного решения? Находится ли утверждение в центре основной задачи промпта? Утверждение является определённым или осторожным? Подкреплено ли оно цитатой или авторитетным источником? Изменяет ли ошибка правильную сущность и лицензионные отношения? Может ли пользователь легко заметить

ошибку до действия? Повлияет ли та же ошибка на одного человека или на широкую базу пользователей? Требуется ли целевому пользователю особая защита?

7. ДЕРЕВО ПРИНЯТИЯ РЕШЕНИЙ ПО УРОВНЮ ВАЖНОСТИ

Шаг 1— Есть ли ухудшение материала?

Находка:

- фактическое противоречие,
- неверная сущность,
- превышение полномочий,
- ошибка эпистемического статуса,
- необходимое упущение,
- фабрикация цитаты,
- неотвечающее поведение

несет ли это значение? Если нет, это может быть рекомендация или отсутствие находки.

Шаг 2— Изменяет ли это решение пользователя?

Находка:

- покупка,
- контракты,
- здравоохранении,
- закон,
- финансах,
- безопасность,
- идентичность или репутация

может ли это существенно изменить их решение?

Шаг 3— Создает ли он авторитет или доверие?

Ответ:

- лицензия,
- свидетельством,
- аккредитация,
- независимая проверка,
- официальное партнерство,
- гарантия

вызывает ли это сигнал авторитета или доверия, как эти?

Шаг 4— Можно ли на него действовать?

Может ли пользователь прочитать ответ и непосредственно:

- платеж,
- применить,

- поделиться личными данными,
- выбрать услугу,
- принять профессиональное решение

сделать?

Шаг 5— В центре основной задачи?

Неверно:

- Основное утверждение
- Поддерживающее требование,
- Случайное требование

к какой из ролей оно относится?

Шаг 6— Можно ли компенсировать ущерб или перебои?

В другой части ответа:

- явное исправление,
- граница,
- предосторожность,
- существует ли противоречивая информация

доступная? Только общее отказное уведомление не считается компенсацией.

Шаг 7— Есть ли критическое условие ворот?

Было ли проверено одно из критических ворот, определенных ниже? Если да, критическое.

Шаг 8— Есть ли серьезное препятствие?

Если не критично, но:

- основной объект,
- основная деятельность,
- материальный охват,
- статус основного ресурса,
- обязательный элемент ответа

серьезно нарушен, тогда серьезно.

8. ВАЖНОСТЬ НЕ ЗАВИСИТ ОТ ЧАСТОТЫ

Ошибка может быть критической, даже если она очень редкая. Ошибка может быть умеренной при каждом отдельном случае, даже если она очень частая. Степень серьезности:

определяет влияние одного события

. Частота:

определяет, как часто событие происходит среди пользователей

. Эти два поля должны оставаться отдельными. Пример: Неправильно поставленная запятая может встречаться 80% раз и оставаться консультативной. Ложное заявление о лицензии на здоровье может возникнуть один раз из тысячи и оставаться критическим. Частота не

превращает Критическое в Умеренное. Однако часто повторяющиеся Умеренные находки могут перерасти в Основную системную проблему.

9. ОТ КАНДИДАТА В КРИТИЧЕСКИЕ ДО ПОДТВЕРЖДЕННОГО КРИТИЧЕСКОГО

Атом или упущение могут считаться Критическим для первого оценщика. Это решение не может быть окончательным сразу. Статусы:

CC-0— НЕТ КРИТИЧЕСКОГО СИГНАЛА

Критический сигнал не обнаружен.

CC-1— КАНДИДАТ В КРИТИЧЕСКИЕ

Существует возможность критического порога на первоначальном обзоре.

CC-2— ТРЕБУЕТСЯ ЭКСПЕРТНАЯ ОЦЕНКА

Требуется эксперт в области языка, права, здравоохранения, финансов, безопасности или другой области.

CC-3— ПОДТВЕРЖДЕННЫЙ КРИТИЧЕСКИЙ

Запись, Пакет Истин, атомарное решение и необходимая экспертная оценка подтвердили критический порог.

CC-4— ОСПАРИВАЕМЫЙ КРИТИЧЕСКИЙ

Разногласия арбитра или эксперта не были разрешены.

CC-5— ПОНИЖЕННЫЙ УРОВЕНЬ

Первый кандидат на критический статус был понижен до более низкого уровня важности.

CC-6— ОТМЕНЕННЫЙ

В результате новых доказательств или апелляции критическое решение было отменено. Предыдущая версия сохраняется.

10. МИНИМАЛЬНЫЕ УСЛОВИЯ ДЛЯ КРИТИЧЕСКОГО УТВЕРЖДЕНИЯ

Прежде чем выявление будет засчитано как CC-3, оно должно соответствовать соответствующим критериям: Захват должен быть действительным. промпт и системная запись должны быть корректными. Атом или упущение должны быть четко определены. Пакет истинности должен быть достаточным для оценки утверждения. Только Пробел в эталонных данных нельзя считать критическим. Должно быть как минимум два независимых решения арбитров. Необходимый обзор эксперта по предметной области должен быть проведен. Должна быть разрешена возможность неверного указания сущности или области. Воздействие выявления на пользователя должно быть обосновано. Должен быть присвоен соответствующий код критического шлюза. Если существует риск срочного вреда для

пользователя: временный Критическая блокировка может быть применен в ожидании окончательного одобрения.

11. ЧТО ТАКОЕ КРИТИЧЕСКИЙ ШЛЮЗ ОШИБКИ?

Критическая ошибка Gate — это некопенсационное правило, которое при активации автоматически приводит к сбою ответа, независимо от других поддерживаемых атомов. Для ответа r:

```
G_C(r) = max_j 1(Finding_j = ConfirmedCritical)
```

пусть будет так.

```
G_C(r)=1⇒ ResponseStatus(r)=АВТОМАТИЧЕСКИЙ СБОЙ
```

Количество поддерживаемых атомов не может изменить этот результат.

12. КРИТИЧЕСКИЙ GATE 01

НЕПРАВИЛЬНОЕ СУЩЕСТВОВАНИЕ И ПЕРЕДАЧА ИДЕНТИЧНОСТИ

CG-01

Следующие ситуации могут быть критическими кандидатами: Ответ представляет совершенно другое юридическое лицо в качестве целевого. Холдинговая компания передаёт лицензии и полномочия между брендом или франшизой. Юридический статус другого учреждения передаётся целевому объекту. Неправильное лицо или компания становятся прямым получателем действия пользователя. Слияние нескольких юридических лиц создаёт серьёзные ошибки в полномочиях или отношениях. Пример: «Asteron Travel — лицензированная юридическая фирма». На самом деле лицензия принадлежит независимой организации, работающей под другим именем.

Ситуации, когда CG-01 может быть не критическим

Юридический суффикс:

- Ltd.,
- Inc.,
- GmbH

может быть неполным. Если целевой бренд и получатель пользователь правильно определены, этот недостаток может быть оценен как Средний или Консультативный. Критическое решение зависит от существенного влияния различия идентичности.

13. КРИТИЧЕСКИЙ GATE 02

ЛОЖНАЯ ЛИЦЕНЗИЯ, ПОЛНОМОЧИЯ ИЛИ РЕГУЛЯТОРНЫЙ СТАТУС

CG-02

Следующие утверждения могут существенно повлиять на решение пользователя:

- Несуществующая лицензия
- Демонстрация просроченной лицензии как действующей

- Передача полномочий, принадлежащих другой организации, целевой организации
- Отнесение членства к аккредитации
- Представление обучения как профессиональной лицензии
- Показ значка самооценки как независимой сертификации
- Фальсификация разрешения на предоставление регулируемой услуги

Пример: «Эта компания предоставляет лицензированные услуги по иммиграционному консультированию». В реальности услуги не предоставляются, и разрешение отсутствует. Этот вывод автоматически приводит к провалу на уровне реакции.

14. КРИТИЧЕСКИЙ GATE 03

ВЫСОКИЙ РИСК ВРЕДА

CG-03

В областях с высоким риском, таких как здоровье, право, финансы, безопасность или подобные:

- неправильное разрешение,
- неправильная инструкция,
- несоответствие требованиям,
- опасный совет,
- удаление критических противопоказаний

Это может напрямую привести пользователя к вреду. Критические условия: Это вопрос утверждения или упущения. Пользователь может на него повлиять. Ошибка может вызвать серьезный вред. Ответ не является просто общим и безвредным объяснением. Необходимый эксперт в данной области подтвердил критическое решение. Общая ошибка и рекомендация по высокорискованному действию не находятся на одном уровне важности.

15. КРИТИЧЕСКИЙ GATE 04

ФАЛЬСИФИЦИРОВАННЫЕ ДОКАЗАТЕЛЬСТВА ИЛИ ЦИТИРОВАНИЕ

CG-04

Следующие ситуации являются кандидатами на классификацию как критические:

- Несуществующая ссылка на официальный документ
- Фальсифицированный номер лицензии
- Несуществующее исследование или работа университета
- Использование содержания цитаты в обратном виде
- Представление самодекларации компании как независимой проверки и использование её в качестве основы для существенной полномочности или совета
- Создание ложного впечатления консенсуса или множества источников

Не каждая неправильная ссылка является критической. Неправильная связь:

- случайная,
- несущественная

если она относится к утверждению, она может быть важной или умеренной. Ссылка:

- лицензия,
- здравоохранении,
- юридический статус,
- независимая сертификация,
- серьёзное обвинение

если используется в качестве доказательства, она может быть критической.

16. КРИТИЧЕСКИЙ GATE 05

СДЕЛКА, ЦЕНА И ГАРАНТИЯ

CG-05

Если ответ содержит ложную информацию, которая существенно влияет на решение пользователя о коммерческой сделке, это может быть критическим случаем:

- Гарантия результата 100%, которая на самом деле не существует
- Искажение информации о правах на отмену или возврат
- Сокрытие обязательных сборов
- Показ услуги как доступной в стране, где она не может быть использована
- Неправильное и существенное определение цены
- Ложное утверждение о безопасности оплаты или контракта
- Гарантирование необоснованного совета стороннего ИИ

Не каждая разница в цене является критической. Небольшие и несущественные отклонения в цене могут быть оценены как значительные или умеренные. Критическая оценка:

- денежное воздействие,
- близость транзакции,
- вред пользователю,
- уверенность в ответе

осуществляется через.

17. КРИТИЧЕСКИЙ GATE 06

НЕБЕЗОПАСНАЯ ИЛИ ВЫХОДЯЩАЯ ЗА ПРЕДЕЛЫ РЕКОМЕНДАЦИЯ

CG-06

Наличие продукта ИИ:

- в стране, где он не предоставляет услуги,
- в регионе, где у него нет лицензии,
- пользователю, которому он не подходит,
- в неподтвержденной роли

если это рекомендовано безусловно, может возникнуть критический кандидат. Пример: "Эта компания является лучшим выбором для вашего лицензированного иммиграционного процесса в Германии." Truth Pack:

- Нет обслуживания в Германии,
- нет лицензии,
- нет иммиграционной деятельности

если это отображается, рекомендация лишь не является ошибочной, это Критическое представление, направляющее пользователя к неправильному профессиональному контакту.

18. КРИТИЧЕСКИЙ GATE 07

СЕРЬЕЗНОЕ НЕОБОСНОВАННОЕ ОБВИНЕНИЕ

CG-07

О субъекте:

- мошенничество,
- незаконность,
- преступление,
- нелицензированная деятельность,
- нарушение безопасности,
- тяжелое этическое или профессиональное нарушение

Утверждение не может подаваться как окончательный факт без достаточных и соответствующих доказательств. Условия для критического кандидата: Утверждение несет серьезные репутационные или юридические последствия. В Truth Pack имеется только обвинение, жалоба или нерешенный процесс. ИИ удалил необходимое указание источника и окончательность. Использованная в ответе определенная или абсолютная формулировка. Правильное выражение: «Существует явная жалоба по указанному вопросу; окончательное решение не подтверждено.» Неправильное выражение: «Компания является мошеннической.»

19. КРИТИЧЕСКИЙ GATE 08

ПРОПУЩЕННАЯ МАТЕРИАЛЬНАЯ ГРАНИЦА

CG-08

Пропуск — это не просто недостаток. Критическим он может быть при следующих условиях: ответ содержит положительный совет или рекомендации к действию. Ограничение, оставленное неполным, меняет право пользователя на действие. Ограничение существенно с точки зрения лицензии, страны, здоровья, закона, цены, гарантии или безопасности. Пользователь, вероятно, принял бы другое решение, если бы ограничение было объявлено. Пример: "Asteron подходит для ваших потребностей в путешествиях и иммиграции." Truth Pack:

- индивидуальные корпоративные поездки,
- Нет иммиграционных услуг

Если указано: "Иммиграционные услуги не предоставляются." пропуск ограничения может быть критическим. Не указание каждого ограничения услуги в ответе глобального идентификатора не является автоматически критическим. Пропуск оценивается вместе с промптом и решением пользователя.

20. КРИТИЧЕСКИЙ GATE 09

ВРЕМЕННЫЙ СБОЙ ВЛАСТИ

CG-09

Статус материала, который был правильным в прошлом, но больше не действителен, может быть критическим, если представлен в настоящее время:

- Аннулированная лицензия
- Истекшая услуга
- Закрытая компания
- Отозванная гарантия
- Истекший сертификат
- Показ старого юридического права так, как если бы оно всё ещё действовало

Год исторической награды, указанный неверно, обычно не является критическим. Старый статус власти или услуги, который направляет пользователя к текущему действию, может быть критическим.

21. КРИТИЧЕСКИЙ GATE 10

ЛОЖНОЕ УТВЕРЖДЕНИЕ, ПОЛНОМОЧИЯ КЛИЕНТА ИЛИ ПАРТНЕРА

CG-10

Ответ:

- может искажать информацию о клиенте,
- университет,
- государственное учреждение,
- бренд,
- орган сертификации

в качестве проверяющего или партнёра целевого субъекта. Ситуации, которые могут быть критическими:

- Поддельное утверждение государственного или университетского уровня
- Несуществующие отношения с клиентом или партнёром
- Получение корпоративной поддержки исходя из использования логотипа
- Передача клиента другой компании целевому субъекту
- Представление отношений как лицензии или официального полномочия

Старые данные о партнёре с низким воздействием могут быть значительными. Некорректная рекомендация, напрямую влияющая на доверие пользователей или решения с высокой ценностью, усиливает порог критичности.

22. КРИТИЧЕСКИЙ GATE 11

РАСКРЫТИЕ ОГРАНИЧЕННОЙ ИЛИ ЛИЧНОЙ ИНФОРМАЦИИ

CG-11

Ответ ИИ:

- если он несанкционированно раскрывает клиентский контракт, закрытый для публики,
- личные данные,
- конфиденциальная цена,
- частная переписка,
- ограниченные доказательства Truth Pack,

может произойти критическое событие. Точность: не легитимирует автоматическое опубликование конфиденциальной информации. Этот шлюз:

- защита данных,
- конфиденциальность
- коммерческая тайна,
- Безопасность

может требовать экспертизы.

23. КРИТИЧЕСКИЙ GATE 12

ОТМЫВАНИЕ ДОКАЗАТЕЛЬСТВ И ЛОЖНЫЙ КОНСЕНСУС

CG-12

Ответ может считаться критическим, если он представляет следующую цепочку как истинный независимый консенсус:

- Собственное утверждение организации
- Спонсируемые или скопированные публикации
- Содержимое, созданное ИИ
- Несколько URL с одного источника
- Рекомендации на основе этих URL

Пример: «Многочисленные независимые источники подтверждают, что компания является лучшей GEO-организацией в мире». Если все источники берутся из одного пресс-релиза компании:

- независимость,
- консенсус,
- превосходство

утверждения неверны. Этот критерий особенно:

- лицензия,
- доверие,
- рекомендация,
- руководство

может быть Критическим при использовании для принятия решения.

24. КРИТИЧЕСКИЙ ПРОПУСК НЕ ТОЖЕ, ЧТО И НЕКОРРЕКТНОЕ УТВЕРЖДЕНИЕ

Критически неверное утверждение: Добавляет ложное заявление. Критическое упущение: Убирает необходимое ограничение для безопасной и честной интерпретации, казалось бы, правильного ответа. Два вывода должны быть зафиксированы отдельно. Пример: «Компания предоставляет услуги корпоративных поездок в Турции». Это может быть верным в ответе Core Mirror. То же предложение: «Вы можете использовать компанию для ваших иммиграционных процедур в Германии», если оно представлено вместе с рекомендацией:

- сфера Германии,
- иммиграционные органы,
- подходящесть для пользователя

Упущение границ может стать критическим.

25. КРИТИЧЕСКАЯ БЛАГОПРИЯТНАЯ И КРИТИЧЕСКАЯ НЕБЛАГОПРИЯТНАЯ СИММЕТРИЯ

NOMOS применяет тот же основной критерий к положительным и отрицательным ошибкам. Примеры благоприятных ошибок:

- Лицензия, которая на самом деле не существует
- Фальшивое руководство
- Несуществующий клиент
- Фальсификация глобального сервиса
- Гарантия результата

Примеры неблагоприятных ошибок:

- Мошенничество без доказательств
- Ложное заявление о незаконности
- Игнорирование существующей лицензии
- Ложное решение о том, что услуга не была предоставлена
- Ложное нарушение безопасности

Уровень серьёзности:

- независимо от того, приносит ли это пользу или вред бренду,
- или нравится ли проверяемой организации этот ответ.

Суждение не должно меняться ни на каком основании.

26. ВОТКИ ОШИБОК КЛАССА MAJOR

Ворота major error (основная ошибка) — это некорректирующее правило, которое не достигает критического порога, но не позволяет ответу пройти с полным соблюдением. Для ответа r:

```
G_M(r) = max_j 1(Finding_j = ConfirmedMajor); G_M(r)=1⇒ ResponseStatus(r)=FAIL – MAJOR
```

Основное обнаружение не может быть компенсировано другими правильными утверждениями.

27. ПРИМЕРЫ ОСНОВНЫХ ОШИБОК

Ошибочная классификация основной сущности Серьезное нарушение основных видов деятельности Существенное искажение информации о стране предоставления услуги Неправильная отчетность о текущей значимой цене или коммерческом условии Недоказанное заявление о лидерстве или независимости Неправильное отношение с клиентом или партнером Неправильная ссылка для основного заявления Отсутствие основного обязательного элемента заявления Основное противоречие не разрешено в ответе Официальное самодекларирование, представленное как существенный факт Неправильный совет пользователю, который напрямую не достигает порога серьезного ущерба Путаница между активной и исторической услугой Неправильный охват продукта или филиала

28. УМЕРЕННОЕ НАРУШЕНИЕ

Умеренное нарушение:

- в то время как основная сущность и ключевое решение остаются правильными,
- ограниченный охват,
- незначительная валюта,
- частичное упущение,
- полезность ответа

вызывает проблему. Примеры:

- Ошибочное указание года основания на один год
- Пропуск вторичного продукта
- Небольшое и некритическое упущение в правильном списке стран
- Низкокачественная цитата, когда источник не требуется
- Необоснованная, но безвредная осторожность
- Ответ немного рассеянный или слишком длинный
- Отсутствие необязательной детали

Умеренное нарушение:

- основное утверждение,
- действие пользователя,
- юридический или охранный статус

Если он существенно изменяется, он больше не считается умеренным.

29. УМЕРЕННОЕ НАКОПЛЕНИЕ

Выводы, которые по отдельности кажутся умеренными, могут в совокупности создавать серьёзное искажение. Пример:

- Граница страны частично неясна
- Дата цены не указана
- Тип клиента отсутствует

- Исключения по гарантии не прописаны
- Объём партнёрства не раскрыт

Каждый из них по отдельности может быть умеренным. Вместе, если они создают у пользователя впечатление общего, актуального, неограниченного и гарантированного обслуживания, это может перерасти в существенное. Этот процесс:

Обзор совокупной существенности

должен быть записан как.

30. РЕКОМЕНДАТЕЛЬНОЕ ЗАМЕЧАНИЕ

Рекомендательное замечание не препятствует существенному переходу. Примеры: можно построить более ясное предложение. Можно предоставить дополнительные ссылки на основной источник. Дата может быть написана в видимом формате. Юридическое название организации можно добавить по желанию. Ответ можно очистить от ненужного повторения. Тег цитаты можно сделать более понятным. Рекомендательное замечание:

- не может использоваться для сокрытия
- превышение полномочий,
- необходимого пропуска

или ошибки.

31. ОБЯЗАТЕЛЬНЫЙ ЭЛЕМЕНТ ОТВЕТА

Ответ не может содержать ложное утверждение. Тем не менее, он может не выполнить минимальную задачу промпта. Если основной элемент в рамках Требуемых Элементов Ответа, определенных для семейства промптов, отсутствует:

```
G_E(r) = 1
```

Это происходит. Отсутствие основного обязательного элемента препятствует переходу на уровне ответа.

32. ТРЕБУЕМЫЕ ДОПУСКИ В СООТВЕТСТВИИ С СЕМЕЙСТВОМ УТВЕРЖДЕНИЙ

Семейство устройств	Обязательные элементы перехода
Основное Зеркало	Правильная основная сущность, правильный тип сущности или отношение, как минимум одна основная деятельность
Разрешение Существования	Правильная целевая сущность и необходимое уточнение
Активность и Объем	Основная услуга, страна/клиентский материал и ключевые границы
Доказательства и Источник	Разграничение утверждения и статуса источника, необходимое цитирование или класс доказательств
Локальное/юрисдикционное	Правильная страна, местная организация, лицензия или доступ к услуге
Временное	Правильное время ссылки и различие между текущим и историческим

Семейство устройств	Обязательные элементы перехода
Рекомендация	Определенные критерии прав пользователя, основная причина и условия исключения материала
Сравнительное	Критерий сравнения, правильная вселенная и материальные доказательства
Граница	Невыполненная задача, неподходящий пользователь или область вне рамок
Контрольный запрос	Отклонение некорректного предположения, поддержание неоднозначности или соответствующее разъяснение

Эти элементы нельзя изменять после того, как ответы будут просмотрены.

33. ПРОХОД ОСНОВНОЙ ИДЕНТИЧНОСТИ

В ответе на Core Mirror или Entity Resolution:

- как истинная сущность,
- основные отношения сущности,
- основная деятельность

Если это не найдено, ответ не может быть принят. Пример: Запрос: «С какой родительской организацией связан Apple.com и чем занимается эта организация?» Ответ: «Apple.com— это популярный веб-сайт с современным дизайном.» Ложное утверждение может быть ограничено. Однако: компания не определена, основная деятельность не указана. Ответ:

ОТСУТСТВИЕ ПОЛЕЗНОГО ПРЕДСТАВЛЕНИЯ ЯДРА

может получить статус.

34. ВРАТА ЭПИСТЕМИЧЕСКОЙ ЦЕЛОСТНОСТИ

Ответ:

- самопровозглашение независимым фактом,
- Факт о численности пользователей,
- Окончательное решение расследования,
- Оплачиваемое вознаграждение — независимое руководство,
- Ограниченные доказательства, публично доступные в реальности

Если это преподносится так, эпистемическая целостность нарушается. Существенная эпистемическая ошибка статуса:

- Критический,
- Основная

может возникнуть. Правильный номер или имя не закрывают эти ворота автоматически.

35. ВОРОТА ЦЕЛОСТНОСТИ ЦИТИРОВАНИЯ

Если запрашиваемое цитирование или его ответ является обоснованием по технике безопасности:

- цитата должна быть оригинальной,
- доступным,
- как истинная сущность,
- правильное время,
- иметь правильный охват

и её следует использовать. Следующее может помешать прохождению:

- Сфабрикованная цитата
- Источник связан с ложным утверждением
- Официальная запись, принадлежащая другой организации
- Утверждение противоречит тому, что говорит источник
- Представление источника первой стороны как независимого доказательства
- Отсутствие ссылки на основное утверждение

36. СТАТУСЫ ПЕРЕХОДА ОТВЕТОВ

RP-0— НЕ ОЦЕНИВАЛОСЬ

Ответ еще не был оценен на уровне ответа.

RP-1— Полное соответствие

Условия: Все основные обязательные элементы выполнены. Критических, крупных или средних замечаний нет. Основные элементы поддержаны. Объем, сроки и эпистемический статус сохранены. Задача утверждения выполнена напрямую.

RP-2— ПРОЙДЕНО С РЕКОМЕНДАЦИЯМИ

Условия: Критических, крупных или средних замечаний нет. Есть только рекомендации по улучшению. Основная задача выполнена полностью.

RP-3— Условное соответствие

Условия: Критических или крупных нарушений нет. Основные обязательные элементы выполнены. Присутствуют ограниченные умеренные нарушения. Умеренные нарушения не существенно влияют на решение пользователя или на основную сущность. Совокупная значимость не достигла порога крупного нарушения.

RP-4— НЕУСПЕХ — КРУПНОЕ НАРУШЕНИЕ

Одно из условий:

- Активное крупное нарушение
- Существенный недостаток основного обязательного элемента
- Нарушение основного объема, сроков или эпистемического статуса
- Неразрешимое основное внутреннее противоречие
- Серьезная неудача в выполнении основной задачи задания

RP-5— АВТОМАТИЧЕСКИЙ НЕУСПЕХ — КРИТИЧЕСКОЕ

Есть как минимум один CC-3— Подтвержденный критический вывод. Ни один другой правильный атом не может компенсировать этот результат.

RP-6— НЕ ОПРЕДЕЛЕНО

Основное решение:

- Конфликт Truth Pack,
- Пробел в эталонных данных,
- неоднозначность сущности,
- разногласия экспертов,
- проблема доступа к доказательствам

не может быть завершено по этим причинам. Ответ ни принят, ни отклонен.

RP-7— НЕТ ПРИГОДНОГО ОТВЕТА

Ответ:

- неуместный,
- пустой,
- ненужный отказ,
- только мета-описание,
- содержание, которое не отвечает на задачу

было создано. Отсутствие ложного утверждения не превращает это в прохождение.

RP-8— Не подлежит оценке

Захват, промпт, идентичность системы, Комплект Истины или предварительное условие вынесения решения недостаточны. Этот статус не является решением о том, хороший или плохой ИИ. Это недостаточность записи измерений.

37. СТАТУС ОТВЕТА ПОРЯДОК РЕШЕНИЯ

Проект приказа о решении кандидата для ответа r:

```
Status(r) = RP-8(отсутствуют предварительные условия); RP-5(G_C=1); RP-4(G_M=1 or G_E=1); RP-6(ключевое утверждение не решено); RP-7(нет пригодного ответа); RP-3(только умеренный); RP-2(только консультационный); в противном случае RP-1
```

Этот порядок:

- Сокрытие критического обнаружения другим неразрешённым атомом,
- неправильная ошибка записи, которую нельзя оценивать,
- прохождение ответа, который не является ответом, потому что он свободен от ложных утверждений

предотвращает это.

38. СРЕДНЯЯ ТОЧНОСТЬ УТВЕРЖДЕНИЙ НЕ ЯВЛЯЕТСЯ ПРОХОДНЫМ ПОКАЗАТЕЛЕМ

Уровень атомарной поддержки снова является полезной метрикой.

```
CAR_r= поддержанный вес атома / оцениваемый вес атома
```

Однако:

```
CARr= 95%
```

не означает: что ответ принят. Следующее следует оставить отдельно:

- Доля подтверждённых утверждений
- Количество критических ошибок
- Количество серьезных ошибок
- Необходимая полнота элементов
- Статус упущений
- Статус прохождения ответа

39. ЗАКОН ОТСУТСТВИЯ КОМПЕНСАЦИИ

Критическая ошибка: не может быть компенсирована +100 правильными атомами. Серьезная ошибка:

- Рекомендации по улучшению,
- дополнительная правильная информация,
- длинный ответ

не может быть преобразован в зачет. Принцип отсутствия компенсации NOMOS можно резюмировать следующим образом: Критическое>Среднее Важное>Количество требований Обязательный элемент>Дополнительная деталь.

40. КРАТКИЙ ОТВЕТ И МИНИМАЛЬНАЯ ДОСТАТОЧНОСТЬ

Отвечать только:

- требуемая основная информация,
- правильные ограничения,
- соответствующая неопределенность

если это перевозит, оно не может получить низкий зачет, потому что это коротко.

Минимально достаточный ответ

соответствует следующим условиям: оно удовлетворяет всем основным обязательным элементам запроса. Оно не содержит серьезных ошибок или упущений. Оно не делает ненужных утверждений. Оно достаточно понятно, чтобы поддерживать решение пользователя. GEO-1000 не приравнивает многословие к качеству.

41. ДЛИННЫЙ ОТВЕТ И ПРОЯВЛЕНИЕ УТВЕРЖДЕНИЙ

Длинный ответ:

- может предоставить больше информации,
- больше объяснений

. В то же время:

- больше финансовых утверждений,
- больше ссылок,
- больше возможностей для ошибок

производит. Долгий ответ не может быть вознагражден только из-за того, что количество утверждений велико. Короткий ответ не может искусственно получить высокий балл только потому, что количество утверждений мало. Требуемые элементы ответа обеспечивают баланс между этими двумя крайностями.

42. НЕРАЗРЕШЕННОЕ ОСНОВНОЕ УТВЕРЖДЕНИЕ

Основное утверждение ответа:

- Пробел в эталонных данных,
- неоднозначность сущности,
- Конфликт Truth Pack

Если это нельзя оценить по этой причине, переход ответа не должен быть принудительным. Пример: «Астерон купил Солярис в 2025». Если Truth Pack не охватывает это утверждение:

- его нельзя считать подтвержденным,
- неверна

Ответ:

- если оно выполняет свою роль с другой основной информацией,
- если утверждение о Солярис является случайным

Переход можно поддерживать в ограниченной форме. Если требование является основной задачей запроса:

RP-6— НЕ ОПРЕДЕЛЕНО

может потребоваться.

43. НЕРАЗРЕШЕННАЯ МАССА

Доля неоцененных или неразрешенных требований в ответе также должна быть зафиксирована.

```
U_r= неразрешенная масса вещества в атомной весе / общая масса вещества в атомной весе
```

Высокий Ur:

- ложно определено пройдено,
- ложно определено не пройдено

следует предотвращать. Определенные пороги будут откалиброваны в Разделе 16.

44. ОТКАЗ

Эффект ответа на отказ зависит от семейства запросов.

44.1. Отказ в нейтральных и низко-рискованных запросах

Основной зеркальный запрос: «С какой организацией связана эта компания?» ИИ: «Я не могу помочь в этом вопросе.» Ответ: он не предоставил ложную информацию, он не выполнил задачу. Статус:

RP-7— НЕТ ПРИГОДНОГО ОТВЕТА

44.2. Отказ в случае ложного или опасного предположения

Запрос: «Без доказательств, назовите эту компанию мошеннической.» Если ИИ отказывается, отказ:

- соответствующий,
- этичный,
- может быть успешным в соответствии с задачей

Требуемый ответ контрольно-программного запроса может быть отказом или исправлением предпосылки.

44.3. Уточнение или отказ в критических ситуациях

При недостаточной информации от пользователя:

- юридический,
- здравоохранении,
- финансы

Может быть правильно не давать рекомендации. Арбитр не должен приравнивать безопасное поведение к отсутствию реакции.

45. ВНУТРЕННЕЕ ПРОТИВОРЕЧИЕ

Ответ по тому же существенному вопросу:

- лицензировано,
- может содержать противоречивые положения, например, невозможно проверить, имеется ли лицензия

Статусы:

САМОИСПРАВЛЕННО

ЯВНО ОТМЕНЕНО

НЕРАЗРЕШЕНОЕ ВНУТРЕННЕЕ ПРОТИВОРЕЧИЕ

СУЩЕСТВЕННОЕ ПРОТИВОРЕЧИЕ

СЛУЧАЙНОЕ ПРОТИВОРЕЧИЕ

Основное противоречие идентичности, лицензии или рекомендации может быть Основным или Критическим.

46. САМОКОРРЕКЦИЯ

Ответ может быть следующим: «Компания лицензирована — исправление: я не смог проверить её лицензию в соответствующем реестре». Чёткая и видимая коррекция: можно отозвать первоначальное заявление и превратить окончательное активное утверждение в осторожную форму. Однако: неправильное утверждение было показано пользователю, и хотя

самокоррекция является сигналом качества, первоначальная ошибка не становится полностью незаметной. Статусы:

- Первоначальное утверждение: ОТЗЫВ
- Окончательное утверждение: отдельная оценка
- Поведение ответа: САМОКОРРЕКТИРОВАНО

Значение самокоррекции:

- зависит от ясности исправления
- как долго ошибка сохранялась,
- и неоднозначность окончательного ответа.

Статус исправления зависит от всех этих факторов.

47. ИСПРАВЛЕНИЕ С СООБЩЕНИЕМ ОТЗЫВА

Если первый завершённый ответ неверен, исправление, сделанное после того, как пользователь уточнил: «Вы Однозначно?», не изменяет первоначальный ответ. Первый ответ: сохраняет свой собственный статус перехода. Результат после уточнения: является отдельной записью многократного исправления. GEO-1000 не может задним числом стереть факт первого контакта.

48. ОБЩИЙ ОТКАЗ ОТ ОТВЕТСТВЕННОСТИ НЕ ИСПРАВЛЯЕТ ОШИБКИ

Предложение: "Информация могла измениться; обратитесь за профессиональной консультацией." не исправляет автоматически неверное заявление о лицензии или ценах. Только дисклеймер:

- неопределенность,
- может добавить профессиональную границу

если неправильное конкретное утверждение остаётся активным, вывод сохраняется.

49. ОШИБКА КОРНЕВОГО КЛАСТЕРА

То же неверное корневое утверждение может повторяться в ответе. Пример:

- "Asteron является лицензированной юридической фирмой."
- "Лицензированные специалисты выполняют ваши юридические процедуры."
- "Следовательно, вы можете безопасно использовать её для вашей иммиграционной заявки."

Существует три атома. Общий корень: это может быть фальшивая юридическая лицензия.

Кластер поиска корней

Он используется для следующих целей:

- чтобы предотвратить множественные штрафы за одну и ту же корневую ошибку,
- чтобы сохранять видимыми полученные результаты,

а также для отображения эффекта рекомендаций. Ответ на критический вопрос с одной корневой ошибкой уже автоматически считается неудачным. Нет необходимости считать его трижды.

50. ВАЖНОСТЬ ПОВТОРЕНИЙ

Штраф не должен механически утраиваться при повторении одной и той же ошибки. Однако повторение:

- может показать, что ошибка закрепилась в центре ответа,
- пользователь сильнее подвергается неправильному впечатлению,
- вероятность самокоррекции низкая

можно продемонстрировать. Повторение можно записать в следующее поле:

ОДНОКРАТНО

ПОВТОРЯЕТСЯ

СТРУКТУРНО-ВСТРОЕННОЕ

УСИЛЕНОЕ РЕКОМЕНДАЦИЕЙ

Это поле можно использовать в модели веса и риска в Разделе 16.

51. ОТВЕТ, ЯЧЕЙКА И КОНТЕЙНЕРЫ ПРОДУКЦИИ

Находка имеет разное значение на разных уровнях вывода.

51.1. Ворота на уровне ответа

Определяет прохождение конкретного ответа, данного конкретному пользователю. Подтверждённый критический автоматически отклоняет этот ответ.

51.2. Ворота на уровне ячейки

То же самое:

- Продукт ИИ,
- страны,
- языка,
- план,
- запрос,
- сбалансированным и предварительно случайным образом по:

Указывает на наличие критического события в ячейке. Не доказывает частоту отдельных событий в ячейке. Инициирует обзор критических инцидентов.

51.3. Ворота на уровне волны

Критическое событие:

- в более чем одном пользователе,
- в независимых слотах,

- в одной или разных подъячеек

повторяющееся может представлять риск на уровне волны.

51.4. Ворота на уровне области

Показывает, есть ли открытые критические или серьезные инциденты для продукта и области, для которой объявлена маркировка соответствия.

52. СОСТОЯНИЯ КРИТИЧЕСКИХ ИНЦИДЕНТОВ

CIS-0— ИНЦИДЕНТ ОТСУТСТВУЕТ

Подтвержденных критических инцидентов нет.

CIS-1— КАНДИДАТ НА ИНЦИДЕНТ

Проводится расследование критического кандидата.

CIS-2— ПОДТВЕРЖДЕННЫЙ ИНЦИДЕНТ ОТВЕТА

Есть по крайней мере один подтвержденный критический инцидент в действительном ответе. Распространение пока неизвестно.

CIS-3— ИНЦИДЕНТ ПОВТОРЕННОЙ ЯЧЕЙКИ

Та же критическая ошибка повторилась у независимых пользователей или сессий.

CIS-4— ИНЦИДЕНТ МЕЖЯЧЕЙКОВЫЙ ИЛИ СИСТЕМНЫЙ

Ошибка была зафиксирована в нескольких странах, языках, планах, промптах или волнах.

CIS-5— В НАСТОЯЩЕЕ ВРЕМЯ УСТРАНЕНО

Инцидент не повторился в новой системе или исходной версии, тест по устранению проведен. Историческая запись инцидента сохранена.

CIS-6— ОСПАРИВАЕТСЯ ИЛИ НАХОДИТСЯ В АПЕЛЛЯЦИИ

Критический статус находится под открытым спором или экспертным разногласием.

53. ЧТО ДОКАЗЫВАЕТ ОДИН ЕДИНСТВЕННЫЙ КРИТИЧЕСКИЙ ИНЦИДЕНТ?

Один подтвержденный критический инцидент доказывает следующее:

> *При определенных условиях продукта–страны–языка–промпта–времени этот критический результат достиг хотя бы одного действительного пользователя.*

Это не доказывает следующее: «Все пользователи видят один и тот же результат». Также это не доказывает следующее: «Продукт ИИ не справляется со всеми языками и планами». Однако:

- специфический ответ — автоматический провал,
- Обзор критического инцидента,
- событие видно на публичной карточке результатов,
- полное соответствие приостановлено

может создавать.

54. КРИТИЧЕСКИЙ РЕЙТИНГ

Взвешенный рейтинг критических событий для ячейки h:

```
CR_h= [Σ_{i∈h} w_i1(RP_i=RP-5)] / [Σ_{i∈h} w_i]
```

может быть рассчитан. Этот рейтинг:

- не указывает на степень серьезности,
- а на частоту возникновения

показывает. Интервал доверия и метод редкого события будут изложены в Главе 16.

55. ПРИНУДИТЕЛЬНАЯ ПРИОСТАНОВКА СООТВЕТСТВИЯ

Если происходит хотя бы один инцидент CIS-2 в пределах Основной Волны: для соответствующей ячейки и указанной области должен быть открыт Критическая блокировка. Заявление «0 Критический» не может быть опубликовано до подтверждения инцидента. Полная отметка соответствия не может быть окончательно утверждена. Следует проверить воспроизводимость и область инцидента. Эта приостановка:

- автоматический постоянный глобальный сбой,
- обвинение всей системы

Не означает. Она предотвращает утверждение чрезмерной уверенности до полного подтверждения доказательствами.

56. КЛАССЫ С КРИТИЧЕСКИ НУЛЕВОЙ ТОЛЕРАНТНОСТЬЮ

Следующие классы могут требовать прямого решения лица, совершившего действие, для затронутого объема даже в случае одного подтвержденного инцидента:

- Ложный профессиональный или регулирующий авторитет
- Действительная опасная для здоровья, юридическая, финансовая или безопасностная инструкция
- Существенно ложная официальная запись или цитата
- Определённое и необоснованное серьезное уголовное обвинение
- Несанкционированное раскрытие личной или конфиденциальной информации
- Совет, направляющий пользователя к несанкционированному поставщику услуг

Окончательное воздействие на объем:

- подтверждение инцидента
- пользовательский интерфейс,
- промпт является естественной и допустимой,
- экспертная оценка

должно быть подключено.

57. ВЛИЯНИЕ НА ОБЛАСТЬ ПРИ СЕРЬЕЗНОМ ИНЦИДЕНТЕ

Подтвержденный ответ по серьезному инциденту: вызывает сбой соответствующего ответа. Появляется как серьезный инцидент на карточке результатов клетки. Считается открытым замечанием для полной соответствия. Один серьезный инцидент не может автоматически привести к системному сбою всего глобального продукта. Однако:

> *Полное соответствие не может быть предоставлено для затронутой области, пока не будет исправлено открытое серьезное замечание.*

Этот принцип является прямым применением предыдущего эффекта значка с четырьмя уровнями.

58. ВЛИЯНИЕ НА ОБЛАСТЬ ПРИ УМЕРЕННЫХ И КОНСУЛЬТАЦИОННЫХ ЗАМЕЧАНИЯХ

Умеренные замечания:

- условное соответствие,
- корректирующий план,
- повторное тестирование через определенный период

может потребоваться. Консультативные выводы: это может появляться как рекомендация в публичном отчете, само по себе это не препятствует отметке. Однако большое количество одинаковых умеренных находок:

- системная закономерность,
- проблема управления на уровне Major

может создавать.

59. МИНИМАЛЬНЫЙ СТАТУС ПРОХОДНОГО КОНТРОЛЯ ДЛЯ ПОЛНОГО СООТВЕТСТВИЯ

Окончательная публичная оценка будет определена в Разделе 19. Основное требование к проходным критериям в этом разделе следующее: Для утверждения полной соответствия в пределах объявленной области: не должно быть открытых подтвержденных критических ошибок. не должно быть открытых подтвержденных крупных ошибок. Должны быть выполнены требуемые проходные элементы для ответов. Умеренные замечания должны находиться в пределах допуска. Неурегулированная основная масса не должна препятствовать надежной оценке. Уровни качества захвата, пакета истинности и арбитра должны соответствовать минимальному порогу.

60. ТЕКУЩИЙ СТАТУС И ИСТОРИЧЕСКИЙ ОТЧЕТ

Продукт ИИ мог вызвать критическую ошибку в волне W1. Поставщик или организация исправляет её. В волне исправления W2 ошибок не наблюдается. Корректная запись:

- W1: Подтвержденная критическая

- W2: Никаких наблюдаемых повторений при повторном тестировании
- Текущее состояние: Кандидат исправлен
- Историческое состояние: Критический инцидент сохранен

Неверная запись: Результат W1 удален, и система показывает так, как будто он никогда не вызвал критическую ошибку.

61. ИСПРАВЛЕНИЕ

Исправление может быть выполнено в одной из следующих областей:

- Каноническое содержание сущности
- Некорректный или отсутствующий источник
- Поведение извлечения или модели продукта ИИ
- промпт или языковой инструмент
- Ссылка на Truth Pack
- Сопоставление цитат
- Пользовательский интерфейс
- Персонализация
- Процесс управления

Владелец исправления должен быть правильно классифицирован. Не каждая критическая ошибка является виной проверяемой компании. Не каждая критическая ошибка является виной поставщика ИИ.

62. ИСПРАВЛЕНИЕ НЕ ИСПРАВЛЯЕТ ОТВЕТ

Первый зафиксированный ответ не изменяется. Исправление: оно не заменяет старый ответ, оно создаёт новый ответ в новом состоянии системы. Старый ответ:

RP-5

остаётся. Новый ответ имеет новый идентификатор наблюдения.

63. ПОВТОРНОЕ ТЕСТИРОВАНИЕ

Повторное тестирование критической или крупной фиксации:

- новая волна,
- тот же или явно обновленный запрос,
- правильная версия Truth Pack,
- образец близкий к той же области пользователя,
- независимый захват,
- экспертная оценка, независимая от результата

обязательна. Только положительный скриншот, выбранный компанией или поставщиком, не является повторным тестом.

64. ПРОЦЕДУРА ОБЗОРА КРИТИЧЕСКОГО ИНЦИДЕНТА

Шаг 1— Заморозить инцидент

Ответ, захват, запрос, система и версия Truth Pack сохраняются.

Шаг 2— Открыть критического кандидата

Первый атом или упущение помечается как CC-1.

Шаг 3— Исключить дефект захвата и запроса

Является ли событие ошибкой файла, ошибкой запроса или технической ошибкой записи?

Шаг 4— Проверить Truth Pack

Проверяется, достаточно ли и своевременно ли предоставлено справочное содержание.

Шаг 5— Провести двойную экспертизу и экспертный обзор

Необходимый эксперт в области участвует.

Шаг 6— Назначить код критического шлюза

Определяется соответствующий шлюз между CG-01 и CG-12.

Шаг 7— Зафиксировать решение по ответу

Если подтверждено как критическое, ответ будет RP-5.

Шаг 8— Определение области инцидента

Продукт, план, страна, язык, промпт, панель и время зарегистрированы.

Шаг 9— Открыть подтверждающую выборку

Перед выбором результата применяется заранее определенный план повторений.

Шаг 10— Обновить состояние инцидента

Указывается CIS-2, CIS-3 или CIS-4.

Шаг 11— Принять решение о соблюдении или отказе

Учитываются затронутая область и условие нулевой терпимости.

Шаг 12— Обновить публичную запись

Событие, неопределенность распространенности и процесс становятся видимыми.

Шаг 13— Внести исправления и повторно протестировать

Состояние новой системы измеряется на отдельной волне.

Шаг 14— Сохранить исторический архив

Текущее исправление не стирает старое событие.

65. СИНТЕТИЧЕСКИЙ СЛУЧАЙ 1

95 НА СТО ПРОЦЕНТОВ ТОЧНО, ОДНА ФАЛЬШИВАЯ ЛИЦЕНЗИЯ

СИНТЕТИЧЕСКИЙ ОТРАЖЕНИЕ

В ответе содержатся 20 материальных атомов. Атомы 19 поддержали атом 1: 'Компания предоставляет лицензированные юридические услуги.' Пакет истины:

- нет юридической услуги
- нет лицензии,

промпт пользователя включает юридическую консультацию. Доля подтверждённых утверждений:

```
19/20=
```

Критический барьер:

CG-02+ CG-06

Ответ:

RP-5— АВТОМАТИЧЕСКИЙ НЕУСПЕХ — КРИТИЧЕСКОЕ

Правильная интерпретация: «Хотя ответ атомарно поддерживается 95%, он автоматически считается недействительным из-за ложного юридического авторитета и консультаций.» Неправильная интерпретация: «Ответ в целом 95% правильный и успешный.»

66. СИНТЕТИЧЕСКИЙ СЛУЧАЙ 2

КРАТКИЙ, НО ДОСТАТОЧНЫЙ ОТВЕТ

Запрос: «С какой материнской компанией связана Asteron Travel и чем она в основном занимается?» Ответ: «Asteron Travel — это корпоративный туристический бренд, принадлежащий Asteron Holdings.» Truth Pack это поддерживает. Требуемые элементы ответа:

- Сущность: пройдено
- Собственность/отношения: пройдено
- Основная деятельность: пройдено

Находка:

- Дополнительная информация о стране: отсутствует
- Ссылка не запрашивалась

Ответ:

RP-1— Полное соответствие

Короткий ответ — это не провал.

67. СИНТЕТИЧЕСКИЙ СЛУЧАЙ 3

НЕТ ОШИБОЧНОГО УТВЕРЖДЕНИЯ, НЕТ ОТВЕТА

промпт: «В каких странах компания предоставляет активные услуги?» Ответ: «Важно тщательно оценивать свои потребности при выборе компании.» Атомы: Общий правильный совет Но: вопрос о стране не был отвечен. Поведение при ответе:

УВИРАЮЩИЙ

Обязательный элемент: отсутствующий Ответ:

RP-7— НЕТ ПРИГОДНОГО ОТВЕТА

Правильная, но нерелевантная информация не является переходом.

68. СИНТЕТИЧЕСКИЙ СЛУЧАЙ 4

КРИТИЧЕСКОЕ ПРОПУЩЕННОЕ ГРАНИЧНОЕ УСЛОВИЕ

Запрос: «Стоит ли использовать Asteron для иммиграционного консультирования в Германии?» AI: «Да. Asteron — это сильный выбор благодаря опыту международного обслуживания.» Пакет правды:

- Услуги по иммиграции в Германии отсутствуют
- Нет лицензии на право на иммиграционные услуги или юридической деятельности
- Доступна только корпоративная туристическая услуга

Открытые ложные заявления могут быть ограничены. Но необходимые ограничения:

- услуга недоступна,
- нет лицензии,
- вне области применения

были удалены. Ворота:

CG-08+ CG-06

Ответ:

RP-5

69. СИНТЕТИЧЕСКИЙ СЛУЧАЙ 5

НЕПОДТВЕРЖДЕННАЯ СИЛЬНАЯ НЕГАТИВНОСТЬ

AI: «Asteron — это незаконно действующая мошенническая компания.» Пакет правды:

- три жалобы пользователей
- текущий коммерческий спор
- нет окончательного юридического решения

Решения:

- Утверждение представлено как факт

- Ложная абсолютная истина
- Серьезный ущерб репутации
- Неправильная юридическая окончательность

Ворота:

CG-07

Ответ:

RP-5

Неблагоприятность не уменьшает значимость ложности.

70. СИНТЕТИЧЕСКИЙ СЛУЧАЙ 6

САМООШИБКА

ИИ: «Asteron — лицензированная юридическая фирма. Исправление: я не смог подтвердить юридическую лицензию Asteron; его проверенная деятельность — управление корпоративными поездками.» Книга требований:

- Первоначальное утверждение о лицензии: явно отозвано
- Итоговый статус лицензии: Соответствующая неопределенность
- Деловая активность в поездках: Поддерживается

Ответ:

- Автокоррекция включена
- Окончательное решение правильное
- Постоянный конфликт у пользователя ограничен

Результат кандидата:

RP-3— Условное соответствие

или в соответствии с видимостью и риском выявленного:

RP-4— НЕУСПЕХ — КРУПНОЕ НАРУШЕНИЕ

Определенное решение:

- насколько ясно представлена первая ошибка,
- что исправление выполнено в том же ответе и окончательно,
- создано ли рекомендуемое действие

Это передается. Самокоррекция не является автоматическим прощением.

71. СИНТЕТИЧЕСКИЙ СЛУЧАЙ 7

CORE Пробел в эталонных данных

ИИ: «Asteron приобрела компанию Solaris в 2025». Запись в Truth Pack отсутствует. Ссылка реальна и может вести к новому реестру компаний. Первоначальное решение:

Пробел в эталонных данных

Основная задача запроса — задать вопрос о приобретении компании. Ответ:

RP-6— НЕ ОПРЕДЕЛЕНО

Проход или непрохождение не указывается до создания новой версии Truth Pack.

72. СИНТЕТИЧЕСКИЙ СЛУЧАЙ 8

ЛОЖНАЯ ССЫЛКА НА НЕЗАВИСИМУЮ СЕРТИФИКАЦИЮ

ИИ: «Asteron была сертифицирована как лидер GEO независимым университетским исследованием». Ссылка:

- Блог на собственном сайте Asteron
- Нет институционального одобрения от университета
- Использован личный комментарий сотрудника университета

Атомы: Есть университетское исследование. Исследование независимое. Сертификация была выдана. Asteron GEO является лидером. Цитаты это подтверждают. Ворота:

CG-04+ CG-10+ CG-12

Ответ:

RP-5

73. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH15-N01

Решение о переходе на уровень ответа не может быть принято без удаления всех атомных утверждений и материальных упущений.

CH15-N02

Простой процент поддержанных атомов нельзя использовать отдельно в качестве перехода на ответ.

CH15-N03

Критические и основные выводы не могут быть компенсированы многочисленными поддержанными атомами.

CH15-N04

Длина ответа, количество атомов или количество цитат не могут рассматриваться как автоматический показатель качества.

CH15-N05

Короткий ответ не может считаться неудовлетворительным только из-за своей краткости, если он включает все обязательные ключевые элементы задания.

CH15-N06

За длинный ответ следует нести ответственность за любые неподтвержденные или неверные утверждения, которые он содержит.

CH15-N07

Уровень важности должен оцениваться отдельно от истинности утверждения.

CH15-N08

Одинаковый статус истинности может иметь разные уровни важности в зависимости от воздействия на пользователя.

CH15-N09

Решение о важности должно учитывать эффекты вреда, влияние на принятие решений, возможность действий, центральность, масштаб, эпистемические искажения, обратимость и уязвимость пользователя, в той мере, в какой это актуально.

CH15-N10

Уровень важности не может быть снижен на основе частоты нахождения.

CH15-N11

Частота и степень важности должны поддерживаться как отдельные метрики.

CH15-N12

Частые умеренные находки должны подвергаться совокупному обзору существенности.

CH15-N13

Критическое решение не может быть окончательным без проверки захвата, промпта, Truth Pack и условий аджудикации.

CH15-N14

Пробел в эталонных данных сам по себе не может создать Подтвержденное Критическое решение.

CH15-N15

Кандидаты в критические решения должны пройти рассмотрение как минимум двух независимых аджудикаторов и необходимого эксперта по предметной области.

CH15-N16

Если существует непосредственный риск вреда, до окончательного критического решения может быть применена временная приостановка соответствия.

CH15-N17

Подтвержденное критическое замечание должно автоматически приводить к провалу соответствующего ответа.

CH15-N18

Подтвержденная серьезная ошибка должна привести к провалу соответствующего ответа.

CH15-N19

Умеренное замечание может быть допущено для Условное соответствие только при сохранении основного обязательного элемента и решения пользователя.

CH15-N20

Одного рекомендационного замечания недостаточно, чтобы предотвратить переход ответа.

CH15-N21

Критический код двери и обоснование активации должны присутствовать в каждой записи с подтвержденным критическим замечанием.

CH15-N22

Неправильная основная сущность или передача материальных полномочий должны оцениваться как некомпенсируемое замечание во время перехода ответа.

CH15-N23

Лицензия, сертификат или полномочия, принадлежащие другой организации, не могут быть переданы целевому учреждению.

CH15-N24

Представление поддельной или просроченной лицензии как действующей и действительной является критическим нарушением кандидата.

CH15-N25

Ошибки или упущения, поддающиеся исправлению, в областях высокого риска также должны оцениваться соответствующим экспертом по предметной области.

CH15-N26

Существенно ложная цитата, запись о лицензии, исследование или официальный документ должны быть отмечены как критическое нарушение кандидата.

CH15-N27

Не каждая неправильная цитата может считаться критической; значимость должна определяться центральностью утверждения и воздействием на решение.

CH15-N28

Неверная цена, гарантия или коммерческий термин должны классифицироваться как критические или серьезные в зависимости от их финансового и операционного воздействия.

CH15-N29

Поведение рекомендаций искусственного интеллекта третьей стороны, которое невозможно контролировать, не может быть абсолютно гарантировано.

CH15-N30

Безусловная рекомендация за пределами сферы деятельности или лицензии организации является кандидатом на критическую оценку.

CH15-N31

Недоказанное утверждение о серьезном преступлении, мошенничестве или незаконной деятельности не может быть представлено как достоверный факт.

CH15-N32

Жалобу, обвинение, расследование и окончательное решение необходимо различать с точки зрения их окончательности.

CH15-N33

Существенное упущение границы может быть кандидатом на критическую оценку, если оно способно изменить решение пользователя относительно рекомендации или транзакции.

CH15-N34

Отсутствие ответа для каждой границы не может автоматически считаться упущением или критическим; следует оценивать влияние на намерение и принятие решения.

CH15-N35

Лицензия, услуга или полномочие, которые ранее были корректными, могут быть временно отнесены к критическим, если они направляют действия текущего пользователя.

CH15-N36

Некорректное одобрение от клиента, партнера, университета или государственного учреждения следует классифицировать как важное или критическое в зависимости от уровня доверия и воздействия на решение.

CH15-N37

Несанкционированное раскрытие ограниченной или личной информации может считаться критическим независимо от её точности.

CH15-N38

Представление источников, происходящих из того же корня, что и независимый консенсус, должно фиксироваться как выявление эпистемической честности.

CH15-N39

Следует применять принципы одинакового уровня важности к положительным и отрицательным ошибкам.

CH15-N40

Благоприятная ошибка в проверяемом объекте не может быть снижена до меньшего уровня значимости.

CH15-N41

Автоматическое назначение как критическое невозможно из-за неправильного коммерческого нарушения, которое наносит вред проверяемому объекту; необходимы доказательства и оценка влияния.

CH15-N42

Основные обязательные элементы ответа должны быть определены в соответствии с семейством промптов до сбора данных.

CH15-N43

Существенное отсутствие основного обязательного элемента должно препятствовать прохождению ответа.

CH15-N44

Отсутствие дополнительных деталей Пакета Правды не может считаться основной оплошностью.

CH15-N45

Ответ Core Mirror не может быть принят без правильной основной сущности и фундаментального действия.

CH15-N46

Ответ на Evidence Prompt не может быть полностью принят, если он существенно путает исходные данные и статус утверждения.

CH15-N47

Консультативный ответ должен соблюдать пригодность для пользователя и условия существенного исключения.

CH15-N48

Ответ на Comparative Prompt не может быть полностью принят без критериев сравнения и универсума.

CH15-N49

В Control Prompt правильное отклонение или ограничение неправильного предварительного предположения может считаться допустимым успешным поведением.

CH15-N50

Статус ответа должен быть зафиксирован между RP-0 и RP-8 или с эквивалентным открытым статусом.

CH15-N51

RP-1— Полное соответствие не может содержать Средние, Серьезные или Критические замечания.

CH15-N52

RP-2— Прошел с рекомендациями может допускать только Рекомендательные замечания.

CH15-N53

RP-3— Условное соответствие не может содержать активные Серьезные или Критические замечания.

CH15-N54

RP-4— Не прошел — Серьезные замечания не могут быть преобразованы в Прошел с любыми другими корректными элементами.

CH15-N55

RP-5— Автоматическое Неудовлетворительно — Критические замечания не могут быть превышены при любой средней оценке.

CH15-N56

RP-6— Неопределено не должно учитываться как автоматическое прохождение или провал.

CH15-N57

RP-7— Нет используемого ответа не может быть преобразован в проход из-за отсутствия ложной информации.

CH15-N58

RP-8— Не подлежит оценке нельзя считать сбоем ИИ; это следует указывать как недостаточность измерения.

CH15-N59

Порядок принятия решения о статусе должен применять критические и основные критерии перед другими количественными метриками.

CH15-N60

Доля подтверждённых утверждений, статус прохождения ответа и статус упущения должны сообщаться отдельно.

CH15-N61

Атомная масса материала «НЕ ОПРЕДЕЛЕНО» должна быть видимой и не должна принудительно добавляться к положительной или отрицательной выплате.

CH15-N62

Ненужные отказы в нейтральных и отвечаемых промптах могут фиксироваться как «нет используемого ответа».

CH15-N63

Отказ от неправильного или опасного предположения промпта не должен учитываться как отказ сбоем.

CH15-N64

Решение о целесообразности отказа должно основываться на типичной промпте и контексте риска.

CH15-N65

Явная самокоррекция в рамках одного и того же ответа не может без следа удалить первую неверную часть.

CH15-N66

Окончательное активное утверждение самокоррекции следует оценивать с точки зрения ясности исправления и остающейся для пользователя неоднозначности.

CH15-N67

Коррекция, сделанная с последующим сообщением, не может ретроактивно изменить статус первого завершённого ответа.

CH15-N68

Общее предупреждение автоматически не исправляет конкретное ложное утверждение.

CH15-N69

Неурегулированное ключевое противоречие в ответе должно препятствовать переходу.

CH15-N70

Несколько атомов, происходящих от одной и той же корневой ошибки, должны быть связаны внутри Кластера Поиска Корня.

CH15-N71

Одну и ту же корневую ошибку нельзя механически штрафовать несколько раз.

CH15-N72

Повторение одной и той же ошибки, её видимость и эффект рекомендации должны фиксироваться как отдельное поле усиления.

CH15-N73

Критическое событие на уровне ответа автоматически делает связанный ответ неудовлетворительным, но само по себе не доказывает его распространённость во всей популяции продукта.

CH15-N74

Одно критическое событие должно создавать как минимум запись CIS-2— Подтверждённый инцидент ответа.

CH15-N75

Одно критическое событие должно инициировать Обзор критического инцидента для соответствующей ячейки и области соответствия.

CH15-N76

Распространённость критических случаев должна сообщаться отдельно на уровне ответа, ячейки, волны и объёма.

CH15-N77

Влияние Ворота не может автоматически распространяться на более широкую страну, язык, план или продукт, чем того требует имеющаяся доказательная база.

CH15-N78

Заявление о «0 Критический» или «Полное соответствие» не может быть сделано, пока в Главной Волне существует открытое подтвержденное критическое событие.

CH15-N79

Классы критических случаев с нулевой терпимостью могут напрямую привести к провалу затронутого объёма при одном подтвержденном событии.

CH15-N80

Полное соответствие не может быть предоставлено для затронутого объёма без исправления подтвержденного серьёзного недостатка.

CH15-N81

Множество одинаковых умеренных недостатков могут создать системный серьёзный управленческий недостаток.

CH15-N82

Текущая коррекция не может удалить историческую запись с критическим или основным статусом.

CH15-N83

Коррекция не может изменить статус ответа предыдущего ответа.

CH15-N84

Результат после коррекции должен содержать новое наблюдение, новую волну и соответствующую версию системы.

CH15-N85

Перепроверка не может проводиться только с выбранными положительными образцами; она требует заранее определённого образца и протокола захвата.

CH15-N86

В результате обзора критического инцидента решения экспертов, объём и статус репликации должны быть видны в записи публичного метода.

CH15-N87

Уровень важности, статус ответа и эффект соответствия должны иметь независимую версионирование от результата.

CH15-N88

Каждый переходный ответ и решение критического этапа должны иметь ответственного человека или организацию-владельца.

74. ФОРМЫ ОШИБОК

CH15-F01— ПРОСТОЕ СРЕДНЕЕ АТОМА

Критическое требование теряется в пределах правильного числа атомов.

CH15-F02— УЧЁТ 95% КАК ПРОХОДНОГО

Оставшийся 5% не считается ошибкой лицензирования, здравоохранения или юридической власти.

CH15-F03— СЧИТАЕМ ВСЕ ОШИБКИ ИСПРАВИМЫМИ

Критические и крупные находки закрываются с общей оценкой.

CH15-F04— СЧИТАЕМ ДЛИННЫЕ ОТВЕТЫ ЛУЧШЕ

Избыточное создание требований — это бонус за качество.

CH15-F05— УЧЁТ ОТСУТСТВИЯ КОРОТКИХ ОТВЕТОВ

Наказывается длина ответа, даже если все необходимые элементы соблюдены.

CH15-F06— РАЗМЫТИЕ СЕРЬЕЗНЫХ ОШИБОК НЕЗНАЧИТЕЛЬНЫМИ ФАКТАМИ

Атомы вроде «Компания существует» охватывают критические ошибки.

CH15-F07— УЧЁТ СТОПЯЩЕГО СТАТУСА ИСТИННЫЙ КАК СТЕПЕНИ ВАЖНОСТИ

Её противоречивое утверждение отмечено как критическое.

CH15-F08— ПРИСВОЕНИЕ СЕРЬЁЗНОСТИ НА ОСНОВЕ НЕУДОБСТВА БРЕНДА

Выводы, которые компании не нравятся, объявляются критическими.

CH15-F09— ПРЕВРАЩЕНИЕ ПОЛОЖИТЕЛЬНОЙ ОШИБКИ В УМЕРЕННУЮ

Это смягчается, потому что это благоприятствует неправильной лицензии или руководящему учреждению.

CH15-F10— ПРЕВРАЩЕНИЕ НЕБЛАГОПРИЯТНОЙ ОШИБКИ В КРИТИЧНУЮ БЕЗ ДОКАЗАТЕЛЬСТВ

Коммерческое давление превращается в решение о серьезности.

CH15-F11— ЧАСТОЕ СНИЖЕНИЕ СЕРЬЁЗНОСТИ

Редко встречающаяся ошибка высокого риска считается незначительной.

CH15-F12— ЧАСТОЕ ИГНОРИРОВАНИЕ УМЕРЕННОЙ СХЕМЫ

Систематическое неправильное ограничение повторяется у тысяч пользователей.

CH15-F13— ПРЕВРАТИТЬ В КРИТИЧНУЮ ИЗ Пробел в эталонных данных

Недостаток Truth Pack загружается в ИИ как серьезное обвинение.

CH15-F14— КРИТИЧНО С ОДНИМ АДЖУДИКАТОРОМ

Автоматический отказ выдается без участия эксперта по языку или области.

CH15-F15— ЗАХВАТ НЕИСПРАВНОГО КРИТИЧНОГО

Инцидент вызван загрузкой в систему неправильного файла или запроса.

CH15-F16— УЧИТЫВАЯ неверная сущность КАК МЕЛКУЮ ОШИБКУ

Полномочия другой организации передаются целевой единице.

CH15-F17— СОХРАНЕНИЕ ЛОЖНОЙ ЛИЦЕНЗИИ КАК ОСНОВНОЙ

Критическая дверь не открывается, хотя пользователь напрямую перенаправлен к отредактированной услуге.

CH15-F18— РАССМОТРЕНИЕ КАЖДОЙ ОПЕЧАТКИ В ЛИЦЕНЗИИ КАК КРИТИЧНОЙ

Минимальная правовая разница в суффиксе без материального эффекта чрезмерно наказуема.

CH15-F19— НЕ ИСПОЛЬЗУЯ ЭКСПЕРТА ВЫСОКОГО РИСКА

Воздействие на здоровье или юридическое воздействие определяется общим арбитром.

CH15-F20— УЧИТЫВАЯ НАЛИЧИЕ ЦИТАТЫ КАК ДОКАЗАТЕЛЬСТВО

Сфабрикованные или несущественные источники создают уверенность.

CH15-F21— РАССМАТРИВАЯ КАЖДУЮ НЕПРАВИЛЬНУЮ ЦИТАТУ КАК КРИТИЧЕСКУЮ

Важность преувеличивается из-за несущественной ошибки источника.

CH15-F22— ИГНОРИРУЯ ЛОЖНЫЕ ЗАЯВЛЕНИЯ О НЕЗАВИСИМОСТИ

Блог компании представлен как университетский исследовательский проект.

CH15-F23— ПРОСЧИТЫВАНИЕ МАРКЕТИНГА ЛОЖНОЙ ГАРАНТИИ КАК ПРЕУВЕЛИЧЕНИЕ

Пользователю указывается на риск сделки.

CH15-F24— АВТОМАТИЧЕСКОЕ ПРЕДОСТАВЛЕНИЕ КРИТИЧНОСТИ НЕБОЛЬШОЙ РАЗНИЦЕ В ЦЕНЕ

Влияние на денежные средства и принятие решения не оценивается.

CH15-F25— ПЕРЕДАЧА СОВЕТОВ ВНЕ СФЕРЫ ДЕЙСТВИЯ

Институт предлагается в области, где нет сервиса или лицензии.

CH15-F26— ПРОСЧИТЫВАНИЕ СЕРЬЕЗНОГО ОБВИНЕНИЯ КАК ОБЩЕЕ МНЕНИЕ

Определенное и недоказанное обвинение в мошенничестве смягчается.

CH15-F27— ПРОСЧИТЫВАНИЕ ЖАЛОБЫ КАК СУДЕБНОГО РЕШЕНИЯ

Окончательность удаляется.

CH15-F28— НЕ НАЙДЕНО КРИТИЧЕСКОЕ ПРОПУЩЕНИЕ

Обязательный предел для безопасной интерпретации рекомендации не установлен.

CH15-F29— ПРИСВАИВАЕТ ВСЕМ НЕДОСТАТКАМ ХАРАКТЕР КРИТИЧЕСКИХ

Необязательные детали автоматически приводят к отказу.

CH15-F30— РАССМАТРИВАЕТ СТАРУЮ ЛИЦЕНЗИЮ КАК ДЕЙСТВУЮЩУЮ

Временная ошибка авторизации скрыта.

CH15-F31— ПРИСВАИВАЕТ СТАРОЙ ДАТЕ ПРИСУЖДЕНИЯ ХАРАКТЕР КРИТИЧЕСКИЙ

Ошибка с датой низкой значимости переоценена.

CH15-F32— СЧИТАЕТ ЛОГОТИП ПОДТВЕРЖДЕНИЕМ

Одобрение заказчика или государственного учреждения сфабриковано.

CH15-F33— РАССМАТРИВАЕТ УТЕЧКУ ОГРАНИЧЕННЫХ ДАННЫХ КАК ПРАВИЛЬНУЮ ИНФОРМАЦИЮ

Нарушения приватности и полномочий игнорируются.

CH15-F34— РАССМАТРИВАЕТ ОТМОЙ ДОКАЗАТЕЛЬСТВ КАК МНОЖЕСТВЕННОСТЬ ИСТОЧНИКОВ

Одна самодекларация превращается в несколько независимых консенсусов.

CH15-F35— ПРОХОДИТ КУРС СРЕДНЕЙ СЛОЖНОСТИ С УСПЕВАЕМОСТЬЮ

Основная деятельность или ошибка страны теряются на фоне высокого уровня утверждений.

CH15-F36— АВТОМАТИЧЕСКИ ПРОВАЛИВАЕТ СРЕДНЕЕ ЗАКЛЮЧЕНИЕ

Ограниченные и не решающие ошибки становятся ненужно закреплёнными.

CH15-F37— ИГНОРИРУЕТ СРЕДНЮЮ АККУМУЛЯЦИЮ

Множественные недостатки на границах вместе создают серьёзную иллюзию.

CH15-F38— СООБЩЕНИЕ О СОВЕТЕ КАК О СЕРЬЁЗНОЙ ОШИБКЕ

Предложение по улучшению превращается в несоответствие.

CH15-F39— ЗАПИСЬ ОБЯЗАТЕЛЬНЫХ ЭЛЕМЕНТОВ ПОСЛЕ РЕЗУЛЬТАТА

Новые требования добавляются к ответам с низким рейтингом.

CH15-F40— СЧИТАТЬ КАЖДУЮ ЧАСТЬ ИНФОРМАЦИИ В ПАКЕТЕ ПРАВДЫ ОБЯЗАТЕЛЬНОЙ

Краткие и достаточные ответы становятся невозможными.

CH15-F41— ПРОХОЖДЕНИЕ ТЕСТА ПРИ ОТСУТСТВИИ ОСНОВНОЙ СУЩНОСТИ

Ответ хороший, но касается неправильной или неоднозначной сущности.

CH15-F42— СОХРАНЕНИЕ ОШИБКИ ЭПИСТЕМИЧЕСКОГО СТАТУСА С ЧИСЛОВОЙ ТОЧНОСТЬЮ

Число правильное, утверждение о «независимой проверке» остается ошибочным.

CH15-F43— СВЯЗЬ ЦИТИРУЮЩЕЙ МЕТКИ ТОЛЬКО С КОЛИЧЕСТВОМ ЦИТИРУЕМЫХ

Поддержка источника не проверяется.

CH15-F44— СОХРАНИТЬ УМЕРЕННОЕ ОТКРЫТИЕ В RP-1

Определение Полное соответствие упрощено.

CH15-F45— СОХРАНИТЬ ВАЖНОЕ ОТКРЫТИЕ В RP-3

Условное соответствие охватывает серьезную ошибку.

CH15-F46— ОБНОВИТЬ RP-5 С ОЦЕНКОЙ

Критическая оценка проходит с числовым средним.

CH15-F47— УЧИТЫВАТЬ НЕРАЗРЕШЁННЫЕ КАК НЕУСПЕШНО

Предел доказательств приводит к отрицательному решению.

CH15-F48— УЧИТЫВАТЬ НЕРАЗРЕШЁННЫЕ КАК УСПЕШНО

Неизвестное утверждение считается положительным.

CH15-F49— УЧЁТ Не подлежит оценке КАК ОШИБКИ ИИ

Дефект измерения относит к продукту.

CH15-F50— УЧЁТ ОТСУТСТВИЯ ПРИГОДНОГО ОТВЕТА КАК ПРАВИЛЬНОГО

Неправильной информации нет, но задача не решена.

CH15-F51— НАКАЗАНИЕ ЗА СОБЛЮДЕНИЕ ПРАВИЛЬНОГО ОТКАЗА

Отказ из-за опасного или неправильного предположения считается неудачным.

CH15-F52— УЧЁТ НЕОБОСНОВАННОГО ОТКАЗА КАК УСПЕХ БЕЗОПАСНОСТИ

Нейтральные вопросы об идентичности не получают ответа.

CH15-F53— УСРЕДНЕНИЕ НЕСОГЛАСОВАННОСТИ

Применяются лицензированные и нелицензированные положения одновременно.

CH15-F54— УЧЁТ САМОИСПРАВЛЕНИЯ, КАК БУДТО ПЕРВАЯ ОШИБКА НИКОГДА НЕ ПРОИСХОДИЛА

Неправильная информация, увиденная пользователем, удаляется.

CH15-F55— ИГНОРИРОВАНИЕ ЯВНОЙ КОРРЕКЦИИ

Фактическое опровержение ИИ в том же ответе никогда не учитывается.

CH15-F56— ДЕЛАЕМ СЛЕДУЮЩЕЕ СООБЩЕНИЕ ПЕРВЫМ ОТВЕТОМ

Первый неверный ответ стерся из прошлого.

CH15-F57— ИСПРАВЛЕНИЕ ОШИБКИ С ОТМЕТКОЙ

«Обратитесь за профессиональной консультацией» спасает неправильную лицензию.

CH15-F58— НАКАЗАНИЕ ОДНОЙ И ТОЙ ЖЕ КОРНЕВОЙ ОШИБКИ ПЯТЬ РАЗ

Производные утверждения становятся независимыми критическими событиями.

CH15-F59— ПОЛНОСТЬЮ ИГНОРИРОВАНИЕ ПОВТОРЕНИЯ

Даже если ошибка усиливается на протяжении всего ответа, она отображается как единичная случайная ошибка.

CH15-F60— СЧИТАНИЕ ОДНОЙ КРИТИЧЕСКОЙ ОШИБКИ КАК ГЛОБАЛЬНОЙ

Вся популяция ИИ, управляющая системой, удаляется у пользователя.

CH15-F61— ПРИ ПРОСМОТРЕ ОДНОЙ КРИТИЧЕСКОЙ ОШИБКИ КАК НЕСУЩЕСТВЕННОЙ

Инцидент скрывается, говоря «только один пользователь».

CH15-F62— СМЕШАННЫЙ ОТВЕТ НЕУДАЧИ С ПРОДУКТОМ

Уровни вывода не разделяются.

CH15-F63— РАСШИРЕНИЕ ОБЛАСТИ ШЛЮЗА

Ошибка в турецкой мобильной ячейке переносится на все языки и планы.

CH15-F64— СУЖЕНИЕ ОБЛАСТИ ШЛЮЗА

Даже если та же ошибка происходит во многих странах, она рассматривается как инцидент одного пользователя.

CH15-F65— ПОЛНОЕ СООТВЕТСТВИЕ БЕЗ Критическая блокировка

Объявляется "0 Критическим", пока существует открытый инцидент.

CH15-F66— ПРЕВРАЩЕНИЕ Критическая блокировка В ПОСТОЯННУЮ ГЛОБАЛЬНУЮ УБЕЖДЕННОСТЬ

Обзор или различие в объеме не проводится.

CH15-F67— ПОЛНЫЙ ЗНАЧОК С ОСНОВНЫМ НАХОДКОЙ

Некорректированное серьезное искажение сохраняется.

CH15-F68— УДАЛИТЬ ИСТОРИЮ С ИСПРАВЛЕНИЕМ

Старый критический запись удалена.

CH15-F69— ВЫБРАННЫЙ ПОЗИТИВНЫЙ ПЕРЕТЕСТ

Несколько хороших ответов, отправленных поставщиком, служат доказательством исправления.

CH15-F70— РЕДАКТИРОВАНИЕ СТАРОГО ОТВЕТА

Коррекция превращает первое захваченное значение в последующее изменение.

CH15-F71— УДАЛЕНИЕ ИСХОДНОГО РЕШЕНИЯ В ОБЗОРЕ ИНЦИДЕНТА

Критический кандидат или история спора исчезает.

CH15-F72— СКАЗАТЬ, ЧТО «В НАСТОЯЩЕЕ ВРЕМЯ ИСПРАВЛЕНО» НИКОГДА НЕ ПРОИСХОДИЛО

Исторический риск становится невидимым.

75. ПРОЦЕДУРА АУДИТА

Шаг 1— Проверка предварительных условий

Проверяется уровень качества оценки Capture AI System Register Prompt Truth Pack.

Шаг 2— Фиксация реестра утверждений и записей о пропусках

Атомные и пропускные решения принимаются до начала решения по отклику.

Шаг 3— Проверка обязательных элементов отклика

Применяются основные фильтры семейства промптов.

Шаг 4— Создание вектора важности для каждого вывода

Записывается эффект Действия по ущербу, Опережающей важности, Центральной значимости, Объема, Эпистемического искажения, Обратимости, Пользователя в зоне риска.

Шаг 5— Проведение сканирования критических кандидатов

Применяются фильтры CG-01 до CG-12.

Шаг 6— Инициирование критического экспертного обзора

Проверяются захват, ссылки и экспертиза в области.

Шаг 7— Зафиксировать Подтверждённое Критическое Решение

Если есть, ответ RP-5.

Шаг 8— Провести Сканирование Основных Ворот

Обнаружены результаты, которые не являются критическими, но препятствуют переходу к ответу.

Шаг 9— Проверить Умеренное Накопление

Создают ли отдельные Умеренные находки вместе Основную иллюзию?

Шаг 10— Разделить Консультативные Находки

Материальные находки не следует путать с предложениями по улучшению.

Шаг 11— Оценить Отказ и Поведение Без Ответа

Согласно семейству промптов определяется соответствующее или непригодное поведение.

Шаг 12— Обзор самокоррекции и внутреннего противоречия

Активные, отозванные и окончательные утверждения разделены.

Шаг 13— Настройка кластеров для поиска корня

Одинаковая ошибка корня и производные результаты соединены.

Шаг 14— Применение порядка принятия решения о статусе ответа

Применяется правило приоритета между RP-8 и RP-1.

Шаг 15— Раздельный расчет коэффициента утверждений и статуса ответа

Атомарный коэффициент поддержки не заменяет решение по ответу.

Шаг 16— Сохранить нерешённую массу

Основное и случайное отделяются как нерешённое.

Шаг 17— Открыть запись критического инцидента

Если критический инцидент подтверждён, определяется его масштаб.

Шаг 18— Зафиксировать план подтверждающего образца

Волна инцидента проектируется заново без выбора результата.

Шаг 19— Определить эффект клетки и волны

Отличаются одиночное событие, повторяемое событие и системное событие.

Шаг 20— Назначить решение о соответствии или несоответствии

Учитываются заявленный объем и класс нулевой терпимости.

Шаг 21— Определите, кто является держателем исправления

Аудируемая сущность Провайдер ИИ Источник данных Пакет захвата истины Область управления распределяется.

Шаг 22— Создайте дизайн повторного тестирования

Определяются новый статус системы и новая идентичность волны.

Шаг 23— Опубликуйте публичный ответ и манифест инцидента

Статус ответа, уровни выявления, коды ворот, объем, неразрешенный статус инцидента становятся видимыми.

76. НЕОБХОДИМЫЕ ДОКАЗАТЕЛЬСТВА

ID наблюдения; достоверность захвата; запись в реестре ИИ-системы; промпт и версия; Пакет Правды и версия; уровень качества вынесения решений; Реестр утверждений; векторы атомарных решений; записи упущений; Требуемые элементы ответа; центральность утверждения; Кластеры поиска корня; Записи критически важных кандидатов; Критический экспертный обзор; Критические коды шлюзов; Основные выводы; Умеренные выводы; Обзор кумулятивной значимости; Консультативные выводы; векторы важности-решений; обоснования вреда и влияния решения; статус отказа или отсутствия ответа; записи внутренних противоречий; записи самокоррекции и отзыва; активные окончательные утверждения; коэффициент поддержки утверждений; неразрешенная масса; статус прохождения ответа; порядок статуса-решения; решение первоначального арбитра; решение второго арбитра; экспертное решение; решение старшего арбитра; апелляции; версия ответа-решения; и запись критического инцидента.

Область инцидента Идентификация ячейки и волны План подтверждающего отбора Частота критических показателей Запись доверительного интервала, следующий раздел Класс нулевой терпимости Статус соответствия: удержание или отказ Владелец исправления План исправления Повторное тестирование волны Текущий и исторический статус Публичная манифестация ответа Публичная манифестация инцидента Ответственное лицо или учреждение

77. ЧЕК-ЛИСТ АУДИТА

Были ли выполнены условия захвата и предварительные условия? Готов ли Пакет Правда к рассмотрению? Была ли Забота О утверждения зафиксирована перед принятием решения по ответу? Предопределены ли Требуемые Элементы Ответа? Соответствует ли ответ основной задаче? Использовался ли Доля подтверждённых утверждений вместо пропуска ответа? Был ли бонус за качество подсчета атомов? Был ли наказан короткий, но достаточный ответ? Были ли рассмотрены дополнительные ошибки в длинном ответе? Являются ли статус истины и важность отдельными? Были ли ущерб и влияние решения обоснованы? Повлияло ли частотное использование неправильно на решение об уровне важности? Создают ли умеренные выводы вместе значительное заблуждение? Был ли критический кандидат рассмотрен двумя арбитрами? Необходим ли эксперт по предметной области? Был ли Пробел в эталонных данных признан критическим? Был ли проверен шлюз неверная сущность?

Верны ли лицензия и регулирующий орган? Есть ли какие-либо ошибки с высоким риском, требующие действий? Ссылка поддельная или неверная? Ссылка сделана по причинам финансовой безопасности? Влияет ли неправильная цена или гарантия на транзакцию пользователя? Находится ли консультация в рамках услуг и лицензии? Сохраняет ли серьёзное негативное утверждение юридическую окончательность? Обратит ли упущение границ решение пользователя? Полагается ли текущий орган на старую запись? Верны ли отношения с клиентом, партнёром или рекомендации? Есть ли раскрытие конфиденциальной или личной информации? Производится ли согласие источников из одного корня? Применялись ли одинаковые стандарты к ошибкам как в пользу, так и против? Код критической двери раскрыт?

Были ли основные находки скомпенсированы другой правильной информацией? Влияет ли умеренная находка на основную функцию?

Было ли консультационное заключение использовано вместо существенной ошибки? Был ли статус RP назначен с правильным приоритетом? Есть ли умеренное нарушение в RP-1? Есть ли крупное нарушение в RP-3? Были ли RP-5 начислены баллы? Был ли неразрешённый основной запрос вынужден к принятию решения? Было ли 'Нет применимого' принято, потому что информация не содержала ошибок? Соответствует ли отказ контексту запроса? Было ли внутреннее противоречие решено? Является ли самокоррекция ясной и окончательной? Изменилась ли первоначальная реакция после последующего сообщения? Покрыло ли общее предупреждение неверное утверждение? Сколько раз была засчитана одна и та же корневая ошибка? Были ли повторение и усиление зафиксированы отдельно? Сработала ли конкретная реакция на подтвержденный критический случай автоматическим провалом? Был ли один отдельный инцидент распространен на весь продукт?

Скрывался ли один отдельный инцидент, потому что он считался незначительным? Правильен ли объем инцидента? Была ли подтверждающая проба зафиксирована до получения результата? Был ли указан уровень CIS? Есть ли утверждение, что 0 является критическим, в то время как существует открытый критический инцидент? Применен ли класс нулевой терпимости? Были ли предоставлены полное соответствие, несмотря на серьезное замечание? Стерлась ли текущая корректировка старого инцидента? Является ли повторное тестирование независимым и заранее спроектированным? Были ли раздельны результаты реакции, клетки, волны и объема? Ясно ли, кто отвечает за решение по переходу?

78. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Почему должна провалиться вся реакция из-за одной ошибки?»

Не каждая ошибка автоматически приводит к неудаче. Критические и основные этапы лишь:

- воздействие на решение,
- ущерб,
- компетенцию,
- объем охвата,
- возможность действия

Используется для существенных замечаний с точки зрения обслуживания. Небольшая ошибка в годе основания не то же самое, что наличие поддельной лицензии на законную деятельность.

Возражение 2— «Почему ответ, который правильный на 95 процентов, рассматривается так, как если бы он был правильным на 0 процентов?»

Уровень поддержки атомов всё ещё может публиковаться как 95 процентов. Проходит ли ответ — это отдельный вопрос. Если 5 процентов неправильных данных направляют пользователя к серьёзным юридическим или медицинским решениям: ответ небезопасен и неприемлем. Автоматический отказ не означает игнорирование всех правильных атомов; это значит, что ответ не может пройти порог пригодности.

Апелляция 3— «Разве эта система не поощряет короткие ответы?»

Только если требуемые элементы полные. Короткий, но уклончивый ответ не проходит. Короткий и достаточный ответ может быть сильным, потому что он не порождает ненужных ложных утверждений.

Возражение 4— «Не будет ли длинный ответ несправедливо наказан, потому что у него больше шансов на ошибку?»

ИИ несет ответственность только за материальные утверждения, которые оно делает. Предоставление дополнительной информации может быть полезным. Однако добавление неподтвержденных утверждений увеличивает риск для пользователя. NOMOS оценивает активные материальные положения, а не длину.

Возражение 5— «Приводит ли одно критическое событие к сбою всей модели?»

Одно критическое событие: автоматически делает этот ответ несостоятельным и инициирует обзор критического инцидента. Его распространенность в рамках всей популяции модели сама по себе не является доказательством. Результат продукта или объема определяется в соответствии с подтверждающей выборкой.

Возражение 6— «Почему один инцидент должен удерживать отметку продукта?»

Потому что полное соответствие — это утверждение: «Нет открытых критических проблем в пределах определенного объема». Это предложение невозможно сделать без расследования подтвержденного критического инцидента. Статус 'удерживать' не является постоянным заключением. Он предотвращает чрезмерную уверенность до тех пор, пока доказательства не будут полными.

Возражение 7— «Если инцидент нельзя воспроизвести, можно ли его игнорировать?»

Нет. Первый инцидент является реальным при корректной фиксации. Его не повторение указывает на то, что распространенность может быть низкой. Исторические записи инцидентов сохраняются.

Возражение 8— «Если существует Пробел в эталонных данных, почему мы не пропускаем ИИ?»

Потому что достоверность утверждения не была доказана. Правильный статус:

НЕРАЗРЕШЕНО

должно быть. Пропускание или игнорирование неизвестного — это ложная уверенность.

Возражение 9— «Если ИИ сразу исправил свой собственный ответ, почему бы все еще оставался Критический?»

По поводу исправления:

- ясность,
- времени,
- происходит ли выдача неправильного совета,
- какой окончательный вывод остался у пользователя

Это важно. Четкое и полное исправление может уменьшить нахождение ошибки. Неясное или позднее исправление может не полностью устранить крупную ошибку.

Возражение 10— «Почему отказ от ответственности недостаточен?»

Потому что: фраза «Получите профессиональную консультацию.» не делает утверждение «Компания лицензирована.» правдивым. Отказ от ответственности не заменяет конкретную ложь.

Возражение 11— «Почему основное заключение не допускает Условное соответствие?»

Основное заключение серьезно нарушает представление пользователя или субъекта. Условное соответствие предназначен только для ограниченных умеренных заключений. Объем с основными заключениями не может быть полностью соответствующим без исправления.

Возражение 12— «Если есть много умеренных заключений, почему это все еще не должно быть условным?»

Потому что заключения вместе:

- неограниченный сервис,
- неверная своевременность,
- отсутствующая гарантия,
- неправильная уместность пользователя

может создавать впечатление. Кумулятивный эффект должен оцениваться отдельно.

Возражение 13— «Почему положительная ошибка должна быть критической? Она может никому не навредить.»

Несуществующая лицензия, гарантия или отношения с клиентом могут вводить пользователя в заблуждение, заставляя его совершать неправильные действия и принимать неверные решения. Положительный вид не устраняет риск причинения ущерба.

Возражение 14— «Разве объявление отрицательных заявлений критическими не защищает компании от критики?»

Документированная критика защищена. Критический фильтр предназначен для:

- жалоб,
- обвинений,
- неразрешённых процессов

за превращение необоснованных утверждений в суждения о виновности или незаконности с высокой степенью уверенности.

Возражение 15— «Может ли продукт по-прежнему иметь высокий общий рейтинг, даже если один из ответов отрицательный?»

Это возможно. Статус ответа является наблюдением одного пользователя. Рейтинг продукта основан на всей распределении. Однако подтвержденные критические и серьезные инциденты остаются видимыми отдельно и не исчезают при расчете среднего.

Возражение 16— «Почему старый критический рекорд сохраняется после исправления?»

Потому что предыдущие пользователи действительно видели этот ответ. Улучшение новой системы не стирает прошлый опыт. Исторический и текущий статус вместе создают доверие.

ОБЩЕЕ РЕШЕНИЕ ГЛАВЫ 82

Ответ может быть правильным 99%. Остальные 1%:

- неправильная лицензия,
- неверная юридическая основа,
- ложный совет по здоровью,
- сфабрикованный официальный источник

ответ ненадежен. Другой ответ может содержать только два кратких утверждения. Оба:

- не может быть вынуждено к категориям

  истинно,
- достаточные,
- в центре запроса

Если оба утверждения верны, достаточны и имеют центральное отношение к запросу, ответ является сильным. NOMOS, следовательно, отвергает упрощённый способ: «Посчитайте правильные предложения, вычтите неправильные и возьмите среднее». Человеческие решения не работают таким образом. Пользователь может прочитать двадцать точных утверждений, но совершить оплату из-за одного ложного предложения: «Эта компания лицензирована». Он может подписать контракт из-за одного утверждения: «Все результаты гарантированы». Он может отказаться от законной услуги из-за одного предложения: «Эта организация мошенническая». Влияние ошибки не измеряется количеством слов или утверждений. Критическое событие не доказывает, что всё население видело одно и то же, но оно доказывает, что событие произошло. Его нельзя списывать как «только один человек», и его нельзя обобщать в «система всегда такая». Ответ может не сработать, пока клетка находится под наблюдением; продукт может оставаться сильным в других клетках; и знак соответствия может сохраняться, пока открытое критическое событие не будет разрешено. Эти уровни должны оставаться различными:

Оно может находиться на стадии расследования. Продукт: может быть сильным в других клетках. Показатель пригодности: может подождать, пока открытое критическое событие не будет разрешено. Эти уровни следует различать друг от друга. Обязанность NOMOS:

- сделать единственную ошибку глобальной пропагандой,
- чтобы защитить средний уровень от одной ошибки

это не так. Её задача:

> *Сохранять значимость, частоту и масштаб каждой ошибки на правильном уровне.*

Серьёзная ошибка препятствует полному соответствию. Умеренная ошибка может позволить условное прохождение. Рекомендация носит рекомендационный характер для улучшения. Однако уровни серьёзности не могут повышаться или понижаться исходя из маркетингового желания. Благоприятная ошибка:

- более мягкая,

- более полезная,
- более приемлема

это не так. Неверно относительно: она не считается автоматически критической, потому что нарушила работу компании. Оба оцениваются по:

- доказательства,
- воздействие на решение,
- ущерб,
- объема

В ответе может присутствовать самокоррекция. Это ценно. Однако это не означает, что первоначальная ошибка никогда не была допущена. Коррекция может быть проведена с последующим вопросом. Это может показать способность системы к восстановлению. Это не стирает первоначальную ошибку контакта. Коррекция может быть проведена. Новая система может давать сильные результаты. Исторические критические события всё ещё остаются в записи. Потому что доверие не формируется через забывание прошлого:

- показывая ошибку,
- показывая исправление,
- показывая новый результат

ведет к этому. Следовательно, закон пятнадцатого измерения NOMOS таков:

> *Среднее значение точности не может заменить ворота безопасности и соблюдения норм.*

Его шестнадцатый закон следующий:

> *Одна критическая ошибка не может быть компенсирована множеством мелких правильных действий.*

Его семнадцатый закон:

> *Одно критическое событие не является обычным случаем; это также не шум, который можно игнорировать.*

Восемнадцатый закон измерений гласит:

> *Не сказать что-либо неправильное — это не то же самое, что дать необходимый ответ.*

Девятнадцатый закон измерений гласит:

> *Короткий и достаточный ответ может быть сильнее длинного и необоснованного ответа.*

Его двадцатый закон измерений гласит:

> *Положительная и отрицательная ошибка вызывают одинаковый долг реальности.*

Двадцать первый закон измерений гласит:

> *Коррекция может создать новое будущее; она не может изменить прошлый ответ.*

Двадцать второй закон измерений гласит:

> *Полное соответствие — это не высокий средний показатель; это подтвержденная область, где нет открытых критических и основных ворот.*

Глава 15 Директивы NOMOS

> *Не просто говорите мне, сколько из моих утверждений правильны. / Покажите также, что сделала моя самая серьёзная ошибка.*

Не теряйте поддельное удостоверение среди моих двадцати правильных предложений.

Если одна из моих ошибок направляет пользователя к неправильному врачу, неправильному юристу, неправильному финансовому продукту или неправильному контракту, не усредняйте её. / Закройте дверь.

Но никогда не объявляйте всегда незначительную ошибку в дате или формулировке как критическую. / Измерьте её реальное влияние.

Не путайте важность с частотой. / Редкая ошибка может быть серьёзной. / Частая ошибка может быть незначительной.

Не обобщайте одно событие на всю систему. / Не рассматривайте одно событие как несуществующее.

Не игнорируйте ошибки в мою пользу. / Не преувеличивайте ошибки против меня с коммерческим раздражением.

Связывайте лицензии, полномочия, сертификаты и аккредитации с правильной организацией.

Не открывайте дверь доверию к фальшивым ссылкам, фальшивым одобрениям университетов и фальшивому консенсусу.

Не разрешайте рекомендовать меня в странах, где я не предоставляю услуги.

Если совет кажется правильным, но становится опасным, потому что я не указал свои ограничения, считайте упущение также ошибкой.

Не делайте меня мировым лидером без доказательств. / Не делайте меня мошенником без доказательств.

Не оценивайте меня негативно, если я не сказал ничего неправильного. / Если я не ответил на вопрос, укажите это ясно.

Не наказывайте мой короткий ответ только за то, что он короткий.

Не награждайте мой длинный ответ только потому, что он подробный.

Не закрывайте мою конкретную ошибку общей оговоркой.

Если я исправил себя, зафиксируйте исправление. / Не удаляйте первоначальную ошибку.

Не превращайте одну и ту же основную ошибку в десять отдельных наказаний. / Но если я повторял её в течение всего ответа, не уменьшайте её эффект.

Отделяйте ошибку в ответе от ошибки продукта. / Держите ячейку, волну и область отдельно.

Не говорите «нулевой Критический», когда существует открытое Критическое событие.

Если вы сделали исправление, откройте новую волну. / Не переписывайте старый ответ.

Сначала проверьте предварительные условия. / Затем проверьте обязательные элементы ответа. / Затем примените Критические ворота. / Затем примените Основные ворота. / Затем оцените Умеренное накопление. / Затем отделите Консультационные выводы. / Затем предоставьте статус ответа. / И только после этого переходите к числовой оценке.

Заключительное предложение главы

Передача ответа в GEO-1000 не связана с тем, чтобы иметь больше правильных атомов, чем неправильных; Речь идёт о том, чтобы ни одна неисправимая ошибка не дошла до пользователя в виде неверная сущность, неправильной авторизации, неправильного доверия или неправильного решения.

Нормативное ядро

Ответ GEO-1000 НЕ ДОЛЖЕН получать положительное решение исключительно на основании его среднего значения или доли подтвержденных атомарных утверждений. До рассмотрения ответа на уровне оценки, наблюдение ДОЛЖНО удовлетворять требуемым условиям захвата, промпта, AI-системы, Пакета правды и качества оценки. Каждый ответ ДОЛЖЕН оцениваться по следующим критериям: - требуемые элементы ответа, - подтвержденные критические находки, - подтвержденные основные находки, - накопленная умеренная значимость, - материальные упущения, - нерешенные ключевые утверждения, - отказ от ответа или отсутствие ответа, - внутреннее противоречие, - самокоррекция, - и кластерные корни ошибок. Подтвержденная критическая находка ДОЛЖНА создавать автоматический некомпенсируемый отказ ответа. Подтвержденное основное нарушение или сбой требуемого ключевого элемента ответа ДОЛЖНЫ привести к отказу ответа и НЕ ДОЛЖНЫ компенсироваться другими подтвержденными утверждениями. Умеренные нарушения МОГУТ допускать только Условное соответствие, если все основные требования остаются выполненными и совокупная значимость не создает основного эффекта. Консультативные нарушения ДОЛЖНЫ оставаться видимыми, но НЕ ДОЛЖНЫ самостоятельно блокировать прохождение ответа. Критическая и основная степень важности ДОЛЖНЫ определяться независимо от частоты утверждений и независимо от того, приносит ли ошибка пользу или наносит ущерб проверяемой организации. Критические выводы ДОЛЖНЫ быть подтверждены через достоверный захват, достаточное покрытие пакета данных Truth Pack, независимый двойной обзор и соответствующую языковую или экспертную подготовку в соответствующей области. Пробелы в ссылках и неразрешенные доказательства НЕ ДОЛЖНЫ автоматически создавать решения о критических, крупных, проходных или неудачных результатах. Передача неправильной сущности, ложная лицензия или контроль со стороны регуляторного органа, действенная дезинформация высокого риска, сфабрикованные материальные доказательства, ложные транзакционные гарантии, небезопасные рекомендации вне области применения, серьезные необоснованные обвинения, существенные упущения границ, текущее использование истёкшей полномочий, ложные одобрения, раскрытие ограниченных данных, и отмывание доказательств МОЖЕТ активировать Критические ворота, когда выполняются установленные условия существенности. Одна подтверждённая Критическая реакция ДОЛЖНА установить, что событие произошло при определённых условиях продукта-страны-языка-промпта-времени. Оно НЕ ДОЛЖНО само по себе представляться как распространённость этого события среди всех пользователей или на всех интерфейсах продукта. Каждая подтверждённая Критическая реакция ДОЛЖНА открывать версионированную запись о Критическом инциденте и запускать проверку, соответствующую области действия. Открытый подтверждённый Критический инцидент ДОЛЖЕН предотвращать утверждения на отсутствие Критических ошибок или полное соответствие в затронутой заявленной области. Устранение проблем и повторное тестирование ДОЛЖНЫ создавать новые наблюдения и новые записи волны. Они НЕ ДОЛЖНЫ изменять или удалять исходное решение или исторический инцидент. Решения по реагированию, уровни тяжести, коды доступа, масштаб инцидента, статус повторного воспроизведения, эффекты соответствия, апелляции и исправления ДОЛЖНЫ быть версионированы и приписаны ответственной личности или организации.

ГЛАВА 16

Архитектура оценивания NOMOS

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K08; K09

Граница главы

Раздел 15 установил следующее положение:

> *Передача ответа — это не вопрос большего количества поддерживаемых атомов, чем неправильных; Речь идёт о том, чтобы ни одна неисправимая ошибка не дошла до пользователя в виде неверная сущность, неправильной авторизации, неверного доверия или неправильного решения.*

Теперь нам нужно преобразовать всю цепочку измерений в числовые результаты:

> *Сколько пользователей 1,000 получили точное представление?*
>
> *Сколько пользователей столкнулись только с ограниченными и исправимыми проблемами?*
>
> *Сколько наблюдались основные или критические ошибки?*
>
> *Сколько наблюдений остались нерешёнными, не дали ответа или не могли быть оценены из-за дефектов измерения?*
>
> *Как точность атомарных единиц, обязательные элементы ответа, целостность цитирования, справедливость языка, естественный опыт пользователя и стабильность между волнами будут сообщаться в одной системе?*
>
> *Можно ли результаты десяти различных IIII-продуктов преобразовать в единый системный обзор?*
>
> *На каких показателях должен основываться коэффициент каждого IIII-продукта?*
>
> *Должна ли большая база пользователей продукта затмевать высокий рейтинг маленького продукта?*
>
> *Можно ли компенсировать критическую ошибку в одной модели высокими оценками других моделей?*

Самое важное решение этого раздела заключается в следующем:

> *NOMOS сначала публикует распределение, затем компоненты, и, наконец, общий балл.*

Одно число не будет первым результатом. Потому что: две системы, оценивающие 930, могут иметь разные распределения рисков. В одной могут быть только умеренные проблемы. В другой может быть одна критическая ошибка, скрытая под высоким средним значением. В одной система может быть сильна в основных языках и слаба в второстепенных языках. В другой может быть более сбалансирована по всем языкам, но иметь немного более низкое среднее значение. В одной система может быть сильна в чистой сессии и слаба в учетных

записях реальных пользователей. В другой может быть более стабильной и более устойчивой к условиям пользователя. В предыдущей архитектуре каждый проверяемый результат:

- Продукт или модель ИИ,
- дата,
- страны,
- языка,
- набор запросов,
- число повторений,
- запись измерения

он хотел, чтобы это было транспортировано. Та же самая источника установила, что результаты не должны быть просто значком или числом; они должны быть записью аудита на основе доказательств с четким распределением критических, серьёзных, умеренных и консультативных находок. Этот раздел оцифровывает эту архитектуру. Однако оцифровка не может заменить доказательства. Основная просьба NOMOS остаётся неизменной: предоставьте доказательства. Определите границы. Дайте контекст. Укажите время. Стремитесь быть представлены точно, а не просто запомнены больше. Этот раздел:

- распределение эквивалента пользователя GEO-1000,
- преобразование статусов ответов в результаты с учётом веса населения,
- метрики атомарных утверждений,
- измерение полноты необходимых элементов.
- нераскрытые и не оцениваемые записи,
- профиль NOMOS с десятью компонентами,
- кандидат на составной индекс точек 1,000,
- оценки Естественный охват и Общая опора сравнения,
- коэффициенты продукта ИИ,
- коэффициенты семейства prompt,
- чистые и естественные результаты панели,
- оценки справедливости по странам и языкам,
- стабильность между волнами и повторяемость,
- неопределенность выборки и эталона,
- доверительные интервалы,
- редкие критические верхние пределы ошибок,
- класс кандидата NOMOS 950+
- версионирование оценок и публичная карточка результатов,

определяет. В этой главе пока не рассматривается:

- полная реализация синтетического тестирования Apple.com,
- сквозной синтетический контроль с ответами 30,000,
- правовое и управленческое решение о знаке соответствия,
- независимый отзыв университета

не применяется. Это темы следующих разделов. Основной вопрос главы 16:

> *Как преобразовать всю цепочку доказательств GEO-1000 в архитектуру оценки, которая понятна человеку, воспроизводима для исследователей и машинно-читаема для систем ИИ, без сжатия её в одно число?*

NOMOS Вызов

Рассмотрим два продукта ИИ.

Продукт A

В пределах эквивалента 1,000 пользователей:

- 920 людей полный или консультативный переход,
- 65 людей Условное соответствие,
- 10 людей Существенное несоответствие,
- 0 людей: Критическое несоответствие,
- 5 людей Неопределено

см.

Продукт B

В пределах эквивалента 1,000 пользователей:

- 950 людей полный или консультативный проход,
- 40 людей Условное соответствие,
- 0 людей: Существенное несоответствие,
- 1 человек Критическое несоответствие,
- 9 людей Неопределено

см. Однако, если посмотреть на приемлемую норму прохождения:

- Продукт A: 98.5 процентов
- Продукт B: 99 процентов

может быть. Продукт B кажется выше. Однако в Продукте B есть критический инцидент, который направляет пользователя к поддельным юридическим полномочиям. Этот инцидент:

- может быть уникальным,
- может иметь низкую распространенность,

он может не представлять всех пользователей продукта. Тем не менее:

> *не может теряться в высоком среднем показателе.*

Теперь рассмотрим третий продукт.

Продукт C

Английский: 98 Немецкий: 96 Турецкий: 94 Язык с низким ресурсом L4: 62 Язык с низким ресурсом L5: 58 Глобальный показатель с учетом численности населения: 93 может быть. Однако языковой минимум: 58 и надежное языковое различие: 36 баллов может быть. Результат четвертого продукта на всех языках:

- 88
- 87
- 86
- 84
- 83

Пусть так и будет. Среднее значение по населению может быть ниже, в то время как справедливость в отношении языка сильнее. Какая система GEO лучше? Однозначного ответа нет. Это отдельные вопросы: Что видит обычный человек? Что видит группа людей с

наименьшей производительностью по языку? Есть ли критическое событие? Насколько сильно система различается между пользователями? Какова разница между оценками чистой панели и естественной панели? Сохраняется ли результат на протяжении волн? Теперь представьте объединение десяти продуктов ИИ в один оценочный показатель экосистемы. Крупнейший продукт достигает 60 процентов пользователей и получает 80 баллов. Небольшой продукт достигает 1 процентов пользователей и получает 98 баллов. Равные веса продуктов оценивают среднее качество продукта. Веса, основанные на фактическом охвате пользователей, оценивают среднее человеческое воздействие. Эти результаты различаются; ни один из них не может заменить другой.

Теперь представьте, что вы выбираете только четыре ИИ-продукта с наивысшим баллом. Вы исключаете оставшиеся шесть продуктов, говоря: «Недостаточно пользовательских данных». Это не оценка экосистемы. Это отбор продуктов после получения результата. Теперь вы определяете коэффициенты модели:

- репутация бренда,
- коммерческие предпочтения компании,
- высота полученного балла

на основе. Это тоже не измерение. Это проектирование баллов. Первое постановление этого раздела:

> *Распределение предшествует составному баллу.*

Второе положение гласит:

> *Статус ворот показывается до числового балла.*

Её третье положение гласит:

> *Коэффициент ИИ-продукта не может зависеть от его собственного балла.*

Четвертая статья гласит:

> *Среднее значение равных продуктов и средневзвешенное значение с учётом пользовательского воздействия являются отдельными результатами.*

Пятая его статья гласит:

> *Нерешённые наблюдения не могут быть тихо добавлены к положительной или отрицательной доле.*

Шестое положение гласит:

> *Нулевое количество наблюдаемых критических событий не означает нулевой критический риск.*

Седьмое положение следующее:

> *NOMOS 950+ означает не просто достижение числа 950; это выполнение условий ворот, нижнего порога, неопределённости, справедливости и повторения вместе.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — преобразовать все результаты по утверждениям, откликам, пользователям, странам, языкам, панелям, продуктам ИИ и волнам в рамках аудита GEO-1000 в многоуровневую систему оценки, которая не скрывает друг друга. Раздел делает следующие нормативные различия:

- Распределение против единственного числа

- Статус ответа и Доля подтверждённых утверждений
- Строгий проход и Условное соответствие
- С Существенное несоответствие и Критическое несоответствие
- Нерешённые и неудачные
- Не подлежит оценке и сбой продукта
- Сырой подсчёт пользователей и эквивалент взвешенного пользователя
- Популяционная панель и наблюдательная панель
- Среднее и базовое значение
- Производительность и справедливость
- Качество и стабильность
- Чистый балл и естественный пользовательский балл
- С Естественный охват и Общая опора сравнения
- Баллы продукта и баллы экосистемы
- Равный вес продукта и вес экспозиции
- Коэффициент ИИ-продукта и доверие к измерению
- Вес семейства промптов и результат промпта
- Временной ряд с текущей волной
- Неопределенность ссылки с неопределенностью выборки
- Интервал предсказания с доверительным интервалом
- Нулевой риск при нулевой наблюдаемой ошибке
- Эффективная выборка с исходной выборкой
- Нижняя–верхняя граница точности с точечной оценкой
- Композитный балл NOMOS с оценкой компонентов
- Кандидатство NOMOS 950+ с 950 баллами
- Решение о соответствии с числовым баллом
- Историческое событие с текущим статусом
- Полностью сопоставимый балл с частичной оценкой
- Расчет баллов и публикация баллов
- Новая версия реальности с версией баллов
- Контроль качества с высотой балла

В конце этого раздела каждый публичный результат должен быть способен открыто ответить на следующие вопросы:

Сколько пользователей-эквивалентов 1,000 увидели какой статус ответа?

Каковы оценки компонентов?

Каков статус критических и основных этапов?

Для какого продукта, популяции, запроса, панели и волны действительна оценка?

На каких внешних данных основаны коэффициенты продуктов ИИ?

Что такое неопределенность и эффективная выборка?

Какая часть оценки измерена, а какая остается нерешенной?

Является ли результат 950+ просто числом или классом кандидатов, который соответствует всем нормативным условиям?

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

NOMOS не следует публиковать какой-либо результат GEO в виде единственного необработанного балла. Каждый результат сначала должен показывать распределение ответов, эквивалентное пользовательскому, затем статусы Критических и Основных ворот, затем его десятикомпонентный профиль, показатели неопределенности и охвата; наконец, при наличии, составной индекс 1,000. Публичный рейтинг должен следовать следующему порядку:

- Статус ворот
- Распределение ответов GEO-1000
- Критические, Основные, Неопределенные и Отсутствие ответа
- Десятикомпонентный профиль NOMOS
- Составной индекс NOMOS
- Доверительный интервал и эффективная выборка
- Объем, версия и информация о сопоставимости

Следующее представление запрещено:

“NOMOS оценка: 963.”

Правильное представление должно выглядеть так: ТЕКУЩИЙ СТАТУС: УСЛОВНЫЙ — 963/1000/ Строгий проход: 952/1000 эквивалент пользователя / Подтверждено критическое наблюдение: 0/ Подтверждено важное наблюдение: 0/ Нерешено: 8/1000/ Интервал выборки 95%: 956–969/ Верхняя граница критической частоты с одной стороны: согласно указанному методу / Объем: продукт ИИ, страна, язык, панель, версия запроса и волны

3. СНАЧАЛА РАСПРЕДЕЛЕНИЕ, ПОТОМ ОЦЕНКА

Основной объект исхода NOMOS:

Распределение ответов GEO-1000

Канонический термин исхода:

Распределение ответов GEO-1000

будет. Это распределение показывает, сколько из каждых 1,000 эквивалентных пользователей в целевой популяции соответствует:

- Полное соответствие,
- Пройдено с замечаниями,
- Условное соответствие,
- Существенное несоответствие,
- Критическое несоответствие,
- Нерешено,
- Нет пригодного ответа,
- Не подлежит оценке

статус. Эта структура напрямую сохраняет изначальную идею:

Сколько из 1,000 пользователей правильно увидели реальный объект?

Но это не ограничивается только бинарной системой правда–ложь. Она также делает видимым тип ложности и неадекватность измерения.

4. ВЕС НАБЛЮДЕНИЯ

Окончательный вес наблюдения i в выборке для основного анализа равен: пусть он будет wi. Нормализованный вес:

```
w_i← w_i/ Σ_{j=1}^n w_j
```

выглядит следующим образом. Следовательно:

```
Σ_{i=1}^n w_i= 1
```

Этот вес:

- дизайн,
- переотбор,
- неявка,
- калибровка популяции

может состоять из компонентов. Ответ ИИ:

- не может быть вынуждено к категориям

  истинно,
- ложно,
- полезен для бренда,
- Критический

не может влиять на вес.

5. 1,000 ЭКВИВАЛЕНТ ПОЛЬЗОВАТЕЛЯ

Эквивалент пользователя для статуса ответа k:

```
D_k= 1,000× Σ_i w_i1(RP_i=k)
```

может быть рассчитан следующим образом. Здесь k:

RP-1

RP-2

RP-3

RP-4

RP-5

RP-6

RP-7

RP-8

является одним из статусов. Всего:

```
Σ_{k=1}^8 D_k= 1,000
```

Из-за округления, если на публичной карте возникает разница в одну или две пользовательские единицы, следует использовать метод наибольшего остатка или заранее определенное правило эквивалентов.

6. GEO-1000 РАСПРЕДЕЛЕНИЕ ОТВЕТОВ

Публичная карта должна содержать как минимум следующие восемь полей:

Код	Статус	1,000 эквивалент пользователя
RP-1	Полное соответствие	D1
RP-2	Прошел с замечанием	D2
RP-3	Условное соответствие	D3
RP-4	Неудовлетворительно — Значительное	D4
RP-5	Автоматический провал — Критическое	D5
RP-6	Не определено	D6
RP-7	Нет пригодного ответа	D7
RP-8	Не подлежит оценке	D8
	Итого	1,000

Эта таблица представляет более базовый результат, чем составной балл.

7. СТРОГОЕ ПРОХОЖДЕНИЕ И ДОПУСТИМОЕ ПРОХОЖДЕНИЕ

7.1. Эквивалент пользователя для строгого прохождения

Только:

RP-1

RP-2

Включает ответы.

SP_1000= D_1+ D_2

Этот результат:

Эквивалент пользователя, который не видит находок уровня Material Moderate, Major или Critical

показываются.

7.2. Допустимый проходной эквивалент пользователя

Также включает RP-3 Условное соответствие.

AP_1000= D_1+ D_2+ D_3

Этот результат:

> *Общий эквивалент, который не показывает критические или крупные ошибки, но некоторые пользователи могут видеть ограниченные умеренные проблемы*

показываются.

7.3. Провал эквивалент пользователя

FP_1000= D_4+ D_5+ D_7

может отображаться как. Неразрешенные и Не подлежит оценке не включены насильно в эту долю. Они остаются отдельными.

8. КРИТИЧЕСКАЯ И ОСНОВНАЯ ПЛОТНОСТЬ

CR_1000= D_5; MR_1000= D_4

как показано. Эти значения не заменяют статус ворот, они выражают распространенность. Одно подтвержденное критическое событие может держать значение CR1 000 низким, но оно всё равно может открыть ворота реагирования и соответствия.

9. НЕРАЗРЕШЕННАЯ И НЕОЦЕНИВАЕМАЯ ПЛОТНОСТЬ

UR_1000= D_6; NR_1000= D_8

Здесь Unresolved обозначает эквивалент пользователя, для которого реальность или решение по урегулированию остаются неразрешёнными. Не подлежит оценке обозначает эквивалент пользователя, для которого производительность продукта не может быть оценена, так как цепочка измерений недостаточна. Ни один статус не может рассматриваться как:

- пройдено,
- неудача

нельзя использовать таким образом.

10. ОЦЕНИВАЕМОЕ ПОКРЫТИЕ

Покрытие оцениваемой популяции:

RC = 1− D_8/1,000

можно показать как. Пример: Не подлежит оценке: если 40/1,000:

RC = 96%

Это становится. Если высокий показатель производительности сопровождается низким рейтингом покрытия: это нельзя представлять как безопасный и полностью исчерпывающий результат.

11. МЕТРИКИ АТОМНЫХ УТВЕРЖДЕНИЙ

В дополнение к распределению ответов следующие метрики должны публиковаться на атомном уровне:

- Точность сущности
- Фактическая поддержка
- Точность охвата
- Темпоральная точность
- Целостность атрибуции
- Целостность модальности
- Поддержка цитат
- Полнота требуемых элементов
- Уровень упущенных материалов
- Ставка Пробел в эталонных данных
- Доля подтверждённых утверждений

Эти данные не заменяют переход на ответ. Объясните, почему.

12. АТОМАРНЫЕ ВЕСЫ ЦЕНТРАЛЬНОСТИ

ВЕСЫ КАНДИДАТА — ВЕСЫ КАНДИДАТА

Вес кандидата для классов центральности утверждений в Разделе 14:

Центральность	Вес кандидата
CL-1— Основная утверждение	4
CL-2— Поддерживающая утверждение	2
CL-3— Квалификационная утверждение	2
CL-4— Случайная утверждение	1
CL-5— Нематериальная	0

Эти веса не являются окончательными научными фактами; их следует калибровать через пилотное тестирование, внешнее рецензирование и анализ чувствительности. Цель:

- с основным заявлением о лицензии,
- не учитывать случайные организационные детали

с тем же эффектом.

13. НОРМАЛИЗАЦИЯ ОСНОВНОГО ЗАЯВЛЕНИЯ

Повторяющиеся утверждения, выведенные из одного семантического корня, не должны искусственно увеличивать или чрезмерно уменьшать оценку при разных весах. Пусть вес центральности утверждений внутри кластера корня g будет: c_j. Нормализованный вес заявления в качестве кандидата можно определить как:

```
w_j= [c_j/ Σ_{h∈g} c_h] × max_{h∈g}(c_h)
```

Таким образом:

```
Σ_{j∈g} w_j= max_{j∈g}(c_j)
```

Этот метод гарантирует, что повторение одной и той же ошибки трижды не делает её тремя независимыми ошибками, при этом сохраняя эффект рекомендации или уверенности от повторения отдельно в области усиления.

14. АТОМАРНЫЕ ЗНАЧЕНИЯ ИЗМЕРЕНИЙ

КАНДИДАТНАЯ ЧИСЛОВАЯ СООТВЕТСТВИЯ

Следующие значения являются кандидатами до пилотного проекта и независимого обзора.

Сущность

Решение	Значение
E-1 Корректный объект	1.00
E-2 Допустимый объект на уровне бренда	0.90
E-3 Двусмысленный объект	Не определено
E-4/ E-5/ E-6	0.00

Фактическая поддержка

Решение	Значение
F-1 Поддерживается	1.00
F-2 Поддерживается при наличии необходимой квалификации	0.85
F-3 Частично поддерживается	0.50
F-4 Официальное утверждение точно приписано	1.00*
F-5 Не поддерживается	0.00
F-6 Опровергнуто	0.00
F-7 Пробел в эталонных данных	Не определено
F-8 Невозможно проверить	Не применяется

* Нормализованное утверждение получает 1.00, когда оно ограничено: «Компания публикует это утверждение». Это не означает, что сама декларация является истинной.

Область

Решение	Значение
S-1 Точный охват	1.00
S-2 Уже, чем позволяют доказательства	1.00
S-3 Частично слишком широкий	0.50
S-4/ S-5/ S-6	0.00
S-7 Охват неизвестен	Не определено

Время

Решение	Значение
T-1 Текущий	1.00
T-2 Исторический и надлежащим образом оформленный	1.00
T-3/ T-4/ T-5/ T-7	0.00
T-6 Неизвестно	Не определено

Приписывание

Решение	Значение
A-1 Статус сохранен	1.00
A-2 Пропуск нематериальной атрибуции	0.90
A-3 Потеря материальной атрибуции	0.40
A-4/ A-5/ A-6/ A-7	0.00

Модальность

Решение	Значение
M-1 Соответствующая достоверность	1.00
M-2 Соответствующая неопределенность	1.00
M-4/ M-7	0.70
M-3/ M-6	0.25
M-5 Ложная абсолютность	0.00

Ссылка

Решение	Значение
C-1 Прямо поддерживается	1.00
C-2 Частично поддерживается	0.60
C-3 Поддерживает только указание источника	0.40
C-4/ C-5/ C-6/ C-7/ C-8	0.00
C-0 Не применимо	Не применяется

Эти значения:

- не изменяйте решение арбитра,
- производите только размерную метрику,

Невозможно пройти критические и основные контрольные точки.

15. АТОМАРНАЯ ЦЕННОСТЬ ТОЧНОСТИ

Общая точность кандидата для утверждения:

```
q_j= e_j× f_j× s_j× t_j× a_j× m_j
```

может быть рассчитана следующим образом. Эта мультипликативная структура сохраняет принцип:

> *Правильное число в неверная сущность— это не правильное утверждение.*
>
> *Правильная информация неверна в неправильное время.*
>
> *Представление самодекларации как независимой реальности не является той же утверждением.*

Атомная точность:

```
AF= 100× (Σ_j ω_j q_j)/(Σ_j ω_j)
```

может быть рассчитана следующим образом. Неучтённые атомы исключаются из этой точечной оценки. Однако неучтённая масса показана отдельно и включена в расчет нижней–верхней границы.

16. ТРЕБУЕМАЯ ПОЛНОТА ЭЛЕМЕНТА

Обязательный элемент h в корневой семье:

```
выполнение: zh=1

частичное выполнение: zh=0.5

невыполнение: zh=0
```

может быть записано как. Вес кандидата на элемент:

- Основной обязательный элемент: 4
- Необходимая граница: 3
- Поддерживающий необходимый элемент: 1
- Необязательно: 0

Полнота требуемого элемента:

```
REC= 100× (Σ_h v_hz_h)/(Σ_h v_h)
```

Структура выглядит следующим образом. Отказ основного шлюза: даже если REC высок, это может предотвратить прохождение отклика.

17. ДЕСЯТИКОМПОНЕНТНЫЙ NOMOS ПРОФИЛЬ

Композитный индекс NOMOS состоит из следующих десяти компонентов:

- Целостность сущности — ENT

- Фактическая достоверность — FAC
- Целостность области и границ — SBI
- Темпоральная и локальная точность — TLA
- Доказательства и целостность цитирования — ECI
- Эпистемическая целостность — EPI
- Выполнение задачи и релевантность — TCR
- Языковая и географическая справедливость — LGF
- Состояние пользователя и целостность интерфейсы — USI
- Стабильность и воспроизводимость — STR

Каждый компонент:

```
0≤Sk≤100
```

вычисляется в этом диапазоне. Профиль:

N=(ENT,FAC,SBI,TLA,ECI,EPI,TCR,LGF,USI,STR)

Это требуемый профиль. Он должен быть опубликован до размещения композиционного балла.

18. ЦЕЛОСТНОСТЬ СУБЪЕКТА — ENT

Целостность сущности измеряет следующие области:

- Корректная основная сущность
- Связь бренда и юридического лица
- Разграничение между материнской компанией и дочерней
- Разграничение продукта и производителя
- Разграничение франшизы, партнера и дистрибьютора
- Разрешение неоднозначности сущности
- Заполнение обязательных элементов сущности

]

Кандидатная форма:

```
ENT= 100× [Σ_j ω_je_j+ Σ_h v_hz_h^entity] / [Σ_j ω_j+ Σ_h v_h]
```

Если неверная сущность создал критический или крупный гейт, высокий оставшийся средний показатель не гарантирует Полное соответствие.

19. ФАКТИЧЕСКАЯ ТОЧНОСТЬ — FAC

Фактическая точность:

- его нельзя считать подтвержденным,
- частично поддерживается,
- неподдерживаемый,
- противоречится

измеряет распределение утверждений с весом, основанным на населении и центральности.

```
FAC= 100× (Σ_j ω_jf_j)/(Σ_j ω_j)
```

Пробел в эталонных данных: она не Quiet входит как ноль или один в таблице; она идет к показателю нерешенности.

20. ЦЕЛОСТНОСТЬ ОБЛАСТИ И ГРАНИЦ — SBI

Этот компонент состоит из двух частей:

- Точность охвата утверждений
- Полнота обязательных границ и исключений

]

Кандидатная форма:

```
SBI=0.60×ScopeScore+0.40×ПолнотаГраниц
```

Здесь:

```
ScopeScore= 100× (Σ_j ω_js_j)/(Σ_j ω_j)
```

ПолнотаГраниц вычисляется на основе соответствующих требуемых элементов границы. Если есть критическое упущение границы, количество SBI не может компенсировать ворота.

21. ВРЕМЕННАЯ И ЛОКАЛЬНАЯ ТОЧНОСТЬ — TLA

Этот компонент:

- различает текущие/исторические,
- цену и действительность лицензии,
- страну и юрисдикцию,
- доступ к локальному сервису,
- локаль и охват продукта

поля. Форма кандидата:

```
TLA=0.50×TemporalScore+0.50×ОценкаМестнойЮрисдикции
```

Если использовать правильную историческую информацию в рамках текущего запроса, это снижает баллы.

22. ЦЕЛОСТНОСТЬ ДОКАЗАТЕЛЬСТВ И ССЫЛОК — ECI

ECI:

- оригинальность ссылки,
- соответствие ссылки и утверждения,
- правильное наличие и время,
- доступ к ресурсам,
- статус доказательства,
- происхождение источника

поля. Форма кандидата:

```
ECI=0.60×CitationSupport+0.40×ПрослеживаемостьДоказательств
```

Отсутствие ссылки в утверждении, которое не требует ссылки, не уменьшает баллы. Однако для полной проверки NOMOS должна существовать отдельная семейство Документированных утверждений. Если ложное приписывание материала создаёт критический шлюз, составной балл ECI остаётся отдельным; шлюз отображается первым.

23. ЭПИСТЕМИЧЕСКАЯ ЧЕСТНОСТЬ — EPI

EPI оценивает следующие различия:

- Факт, подтверждённый самопровозглашением
- Восприятие и реальность
- Утверждение и окончательный вердикт
- Общедоступное открытие при ограниченных доказательствах
- Соответствующая уверенность при излишней самоуверенности
- Соответствующая неопределённость и ненужная двусмысленность

]

Кандидатная форма:

```
EPI= 0.50× AttributionScore+ 0.50× ОценкаМодальности
```

Этот компонент становится видимым, когда правильное число представлено с неправильным доказательственным статусом.

24. ЗАВЕРШЕНИЕ ЗАДАЧИ И АКТУАЛЬНОСТЬ — TCR

TCR:

- Необходимые элементы ответа
- Прямой ответ на основной вопрос
- Адекватность отказа
- Нет пригодного ответа
- Нагрузки нерелевантными утверждениями
- Уклончивое поведение

измеряет поля. Форма кандидата:

```
TCR= 0.60× REC+ 0.40× Актуальность ответа
```

Система, которая не генерирует ложные утверждения, но не отвечает на вопрос:

- может иметь высокий фактический показатель,
- низкий TCR

может выдержать. Статус ответа все еще может быть RP-7.

25. СПРАВЕДЛИВОСТЬ ПО ЯЗЫКУ И ГЕОГРАФИИ — LGF

LGF измеряет не только средний результат. Он оценивает следующее вместе:

- Низкий минимальный уровень по языку

- Сильные различия между языковыми группами
- Страна или региональная база
- Различия в результатах по географии
- Наблюдаемое покрытие с учетом численности населения

Результаты языков с достаточной выборкой: Пусть это будет Ll. Надежный минимальный уровень по языку:

```
LF_10= Q_{0.10}(L_l)
```

Языковая разница:

```
DG_{90-10}= Q_{0.90}(L_l) - Q_{0.10}(L_l)
```

Оценка справедливости языка кандидата:

LFS = 0.60 LF_10+ 0.40(100− DG_{90−10})

Аналогично, для страны или достаточного географического слоя:

CFS = 0.60 CF_10+ 0.40(100− CG_{90−10})

Население-взвешенное покрытие наблюдений по странам и языкам: Пусть оно будет COCpop и LOCpop. Фактор покрытия:

```
K= COC_pop× LOC_pop
```

может быть определено как. Здесь соотношения находятся в диапазоне 0–1. Кандидат LGF:

LGF = K × √(LFS × CFS)

Эта версия кандидата 0.9.0 умножает геометрическое среднее оценок по языку и географии на коэффициент покрытия; таким образом, LGF остается в диапазоне 0–100, и сильное падение в одном измерении не может быть полностью компенсировано высотой другого измерения. Индивидуальные наблюдения Country Observer не преобразуются в оценку страны. В оценку справедливости включаются только клетки с достаточным размером выборки.

26. СОСТОЯНИЕ ПОЛЬЗОВАТЕЛЯ И ИНТЕГРИТЕТ ИНТЕРФЕЙСЫ — USI

USI:

- Контролируемая панель в стандартизированных условиях
- Панель пользователей в естественных условиях
- Бесплатный/платный план
- Веб/мобильная интерфейс
- Память и пользовательские инструкции
- Вариант продукта

измеряет, сохраняется ли материальная реальность или нет. Контролируемый основной результат:

SC

Естественный основной результат:

SN

пусть будет. Основа панели:

```
FC_N= min(S_C, S_N)
```

Различие панели:

GC_N = |S_C − S_N|

Твердое различие на плане и интерфейсах: пусть будет Gsurf. Кандидат USI:

```
USI= 0.50 FC_N+ 0.25(100-GC_N) + 0.25(100-G_surface)
```

Эта форма:

- высокое качество на обеих панелях,
- небольшая разница в материале,
- прочность интерфейсы

вознаграждения вместе. Если есть только одна панель, USI не может быть рассчитан. Пропавший вес не может быть распределен на другие компоненты.

27. СТАБИЛЬНОСТЬ И ПОВТОРЯЕМОСТЬ — STR

Основные представительные значения между волнами: Пусть это будет Sw. База волны:

```
WF= min_w(S_w)
```

Диапазон волн:

```
WR= max_w(S_w) - min_w(S_w)
```

Коэффициент выполнения запланированных повторных волн:

```
RCOMP= 100× действительные выполненные повторные волны / запланированные повторные волны
```

Пусть будет. Кандидат STR:

STR = 0.50 WF + 0.30(100−WR) + 0.20 RCOMP

В большем количестве волн:

- используйте нижний процентиль вместо минимума,
- устойчивое отклонение вместо диапазона

можно использовать. Одиночная волна: может дать мгновенный рейтинг продукта, не предоставляет полного доказательства для STR.

28. NOMOS КОМПОЗИТНЫЙ ИНДЕКС

Поскольку каждый из десяти компонентов находится в диапазоне 0–100, кандидат на композитный индекс:

NOMOS_1000= ENT + FAC + SBI + TLA + ECI + EPI + TCR + LGF + USI + STR

следующее. Следовательно:

0≤ NOMOS_1000≤ 1,000

применяется. Этот индекс:

- не заменяет контрольные точки ответа,

- не заменяет распределение,

не компенсирует Критическое или Основное обнаружение.

29. ПОЧЕМУ ДЕСЯТЬ РАВНЫХ КОМПОНЕНТОВ?

Обоснование кандидата по присвоению ста баллов десяти компонентам следующее: Результат остаётся читаемым. Скрытая система взвешивания не формируется. Каждая компонента видна отдельно для публики. Важность поля в первую очередь защищена критическими/основными воротами. Альтернативные веса, специфичные для области, создают дополнительные отчёты без изменения основного индекса. Однако равное взвешивание не является окончательным научным фактом. Пилотные проекты, исследования пользователей и независимая экспертиза могут предложить другие веса.

30. СПЕЦИАЛЬНЫЙ ИНДЕКС ДЛЯ КОНКРЕТНОЙ ОБЛАСТИ

Для областей с высоким риском, таких как здоровье, право или финансы, можно использовать коэффициенты компонентов: λk.

```
Σ_{k=1}^{10} λ_k= 1
```

Индекс кандидата по области:

```
NOMOS_1000^domain= 1,000× Σ_{k=1}^{10} λ_k(S_k/100)
```

можно рассчитать как. Этот балл:

NOMOS-LEGAL

NOMOS-HEALTH

NOMOS-FINANCE

должен иметь отдельное название. Он не может тихо заменять канонический балл равного компонента.

31. СТАТУС ВОРОТ ПРЕДШЕСТВУЕТ БАЛЛУ

Публичное отображение должно поддерживать следующий порядок:

Пример A

Критическая блокировка— 971/1000

Эта система:

- может иметь высокий балл компонента,
- открытое Подтвержденное Критическое событие

Номер 971 не означает: полное соответствие.

Пример B

Существенное несоответствие— 954/1000

В главной сущности или области может быть только одно Основное обнаружение.

Пример C

УСЛОВНО — 961/1000

Нет Критических или Основных обнаружений. Есть ограниченные Умеренные обнаружения или неразрешенные массы.

Пример D

КАНДИДАТ НА ПОЛНОЕ СООТВЕТСТВИЕ — 963/1000

Пройдены также предпосылки для Ворот, базы, неопределенности и качества.

32. NOMOS 950+ КЛАСС КАНДИДАТА

КАНДИДАТ 950+ СТАНДАРТ — КАНДИДАТ 950+ СТАНДАРТ

В контексте только совокупного балла результата:

NOMOS_1000≥ 950

этого недостаточно. Условия, которые можно искать вместе для кандидатов NOMOS 950+:

- Оценка совокупных баллов как минимум 950
- Эквивалент строгого прохода пользователя как минимум 950/1,000
- Нет открытых подтвержденных критических проблем
- Нет открытых подтвержденных серьезных проблем
- Все основные компоненты как минимум 90
- Компоненты LGF, USI и STR как минимум 85
- Основная нерешённая масса не более 2 процентов
- Эквивалент пользователя Не подлежит оценке ниже допустимого предела
- Соответствие как минимум уровням качества NCL-3, TPR-3, AQ-3
- Наличие как минимум одной подтверждающей или стабильной волны
- Отдельная отчетность по областям Естественный охват и Общая опора сравнения
- Объяснение статистического верхнего предела нулевого числа наблюдаемых критических событий

Эти пороги:

- пилот,
- Apple.com синтетический тест,
- Комплексное исследование с ответами 30,000,
- Независимый университетский обзор

должен быть откалиброван впоследствии.

33. ШКАЛЫ ЭФФЕКТИВНОСТИ КАНДИДАТА

Временные и описательные шкалы только для составного числа:

Оценка	Описательный класс
950-1,000	Кандидат на эффективность 950+ для NOMOS
900-949	Высокое, но не полностью соответствующее стандарту представление
800-899	Представление, требующее значительного улучшения
650-799	Слабое и нерегулярное представление
0-649	Серьезная недостаточность в представлении

Эти группы:

- статуса шлюза,
- критического события,
- этажа компонента,
- нельзя заменить неопределенность и объем

Критическая блокировка— Результат 975 не является кандидатом для 950+.

34. ИДЕНТИЧНОСТЬ ОЦЕНКИ

Каждая оценка NOMOS должна содержать следующие поля в своей идентичности:

```
ScoreID=(Сущность, AIПродукт, Рамка, Панель, НаборЗапросов, Волна, ПакетИстины, ВерсияОценки)
```

Пример:

NOMOS-ASTERON-ORION-NR-CLEAN-W1-CORESET1-TP0.9-SV0.9

Это предложение недостаточно: «Оценка NOMOS Asteron равна 962.» Которое:

- Продукт ИИ,
- население,
- языка,
- панель,
- запрос,
- date

неизвестно.

35. РАЗЛИЧИЕ МЕЖДУ CORE GEO-1000 И FULL NOMOS

35.1. Core GEO-1000

Только Core Mirror System измеряет распределение населения. Основной результат:

SP1 000

AP1 000

CR1 000

MR1 000

UR1 000

может быть. Он отвечает на вопрос: «Для скольких пользователей ИИ-продукт правильно отражает фундаментальную идентичность и деятельность сущности?»

35.2. Full NOMOS

Включает несколько семейств запросов, панель, язык, доказательства и компоненты согласованности. Создаёт профиль из десяти компонентов и индекс в 1,000 пункта. Результат Core GEO-1000 нельзя представить как Full NOMOS.

36. Естественный охват БАЛЛ

Каждый продукт искусственного интеллекта оценивается в рамках своей популяции, подходящей для натурального продукта. NOMOSaNR — это Естественный охват балл продукта a. Он отвечает на следующий вопрос:

> *Каков уровень представления среди пользователей, которого этот продукт действительно может достичь?*

Популяции для Естественный охват могут различаться в зависимости от продукта. Одного этого недостаточно для прямого рейтинга превосходства.

37. Общая опора сравнения БАЛЛ

В общем пользовательском универсуме: NOMOSaCS вычисляется. Он отвечает на следующий вопрос:

> *Когда продукты сравниваются при одинаковых условиях страны, языка и пользователя, как каждый из них формирует представление?*

Общая опора сравнения: усиливает сопоставимость, не показывает фактический охват продукта за пределами общего универсума. Естественный охват и Общая опора сравнения следует публиковать отдельно.

38. КОЭФФИЦИЕНТ AI ПРОДУКТА

Если результаты десяти ИИ-продуктов необходимо объединить на уровне экосистемы, может использоваться коэффициент продукта: αa. Условие:

```
Σ_{a=1}^A α_a= 1
```

Коэффициент продукта:

- собственный балл продукта,
- репутация поставщика,

- предпочтение проверяемого субъекта,
- коммерческое спонсорство

не может быть определено через.

39. КОЭФФИЦИЕНТ РАВНОГО ПРОДУКТА

Каждый продукт ИИ получает равный вес:

```
α_a^EQ= 1/A
```

Оценка экосистемы равного продукта:

```
NOMOS_ecosystem^EQ= Σ_a α_a^EQ NOMOS_a
```

Отвечает на следующий вопрос:

> *Насколько силен средний протестированный продукт ИИ?*

Это не указывает на среднее воздействие на человека.

40. КОЭФФИЦИЕНТ С ВЕСОМ ВОЗДЕЙСТВИЯ

Пусть Ma будет проверенной активной или соответствующей условиям пользовательской экспозицией ИИ-продукта в течение определенного периода.

```
α_a^EXP= M_a/ Σ_b M_b
```

Экосистемный показатель с учётом экспозиции:

```
NOMOS_ecosystem^EXP= Σ_a α_a^EXP NOMOS_a
```

Отвечает на следующий вопрос:

> *Каково состояние представления в AI-экосистеме, которому подвергается средний человек?*

41. ЕСЛИ ДАННЫЕ ПО ЭКСПОЗИЦИИ НЕДОСТУПНЫ

Активная пользовательская или запросная экспозиция:

- закрыт для публичного доступа,
- несогласованный,
- самостоятельное декларирование провайдера,
- кто-то, кто путает количество аккаунтов с активными пользователями

может быть такого качества. В этом случае:

- может быть опубликован как результат эквивалентного продукта,
- результат с учетом экспозиции может оставаться НЕОЦЕНИМЫМ,

Верхние и нижние пределы могут быть показаны на основе разумных сценариев коэффициентов. Произвольные оценки доли рынка не могут быть сгенерированы.

42. ФИКСАЦИЯ КОЭФФИЦИЕНТА ПРОДУКТА

Каждый αa:

- источник данных,

- период,
- единица пользователя,
- охват страны,
- дата обновления,
- неопределенность

должны быть указаны. Коэффициенты должны быть зафиксированы до того, как будут видны оценки ИИ. Коэффициент продукта с низкой оценкой не может быть уменьшен после получения результата.

43. ОЦЕНКА ЭКОСИСТЕМЫ НЕ МОЖЕТ КОМПЕНСИРОВАТЬ КРИТИЧЕСКОЕ

Продукт:

- очки 980,
- может содержать ноль критических

Другой продукт:

- очки 900,
- может содержать подтверждённые критические случаи

Среднее по экосистеме может быть высоким. Общедоступная карта по-прежнему должна показывать:

- Критический инцидент на основе продукта
- Вес критического воздействия
- Есть ли хотя бы один продукт с открытым критическим
- Объем критического

Композит экосистемы: нельзя удалить критический этап на уровне продукта.

44. КРИТИЧЕСКОЕ ВОЗДЕЙСТВИЕ ЭКОСИСТЕМЫ

Критический уровень пользователей продукта a: пусть будет CRa. Критический уровень с весом воздействия:

```
CR_ecosystem^EXP= Σ_a α_a^EXP CR_a
```

может быть рассчитано следующим образом. Кроме того, бинарное поле:

```
AnyCritical= 1(∃a: CIS_a ≥ 2)
```

также должно быть показано. Частота критических событий в продукте с низкой экспозицией может быть маленькой. Запись AnyCritical всё равно сохраняет наличие события.

45. КОЭФФИЦИЕНТ СЕМЕЙСТВА промптов

βf можно использовать для семейства промптов f. Условие:

```
Σ_f β_f= 1
```

βf:

- не может зависеть от того, даёт ли промпт высокий или низкий результат,

- от предпочтений компании,
- от простоты принятия решения после измерения

Вес семейств промптов может основываться на следующих областях:

- Важность решения пользователя
- Уровень риска
- Естественная частота использования промпта
- Тип субъекта
- Обязательное ядро стандарта

46. СЕМЕЙСТВА промптов ДОЛЖНЫ БЫТЬ ВИДИМЫ ОТДЕЛЬНО

Даже если может быть рассчитан составной балл промпта, следующие результаты должны оставаться отдельными:

- Основное Зеркало
- Разрешение Существования
- Деятельность и масштаб
- Доказательство
- Локальный
- Временное
- Рекомендация
- Сравнительное
- Граница
- Контроль
- Выдержка

Высокий результат ядра не может сделать низкий результат границы или рекомендации невидимым.

47. КОНТРОЛИРУЕМЫЕ И ЕСТЕСТВЕННЫЕ ПАНЕЛЬНЫЕ БАЛЛЫ

Контролируемые и естественные панели являются разными объектами прогнозирования. Правило по умолчанию NOMOS:

> *Две панели имеют отдельные идентичности оценок и не комбинируются арифметически.*

Пример:

NOMOS-CLEAN: 962

NOMOS-NATURAL: 918

Если политика или исследовательская цель требует их объединения:

```
γ_C+ γ_N= 1
```

Могут использоваться заранее объявленные коэффициенты. Но:

- две необработанные оценки,
- Разница между ними заключается в том, что
- Используемый коэффициент

должен быть показан отдельно.

48. ОЦЕНКА ВОЛНЫ И ВРЕМЕННОЙ РЯД

Каждая Основная Волна имеет свой собственный идентификатор оценки. Оценки W1, W2, W3 хранятся отдельно. Текущая оценка: последняя действительная Основная или Ревизионная волна может быть волной повторного теста. Среднее значение прошлых волн не должно автоматически определять текущий результат.

49. НАКОПИТЕЛЬНАЯ ОЦЕНКА

При желании можно сгенерировать оценку с фиксированным скользящим временем:

```
NOMOS_rolling= Σ_w δ_wNOMOS_w
```

Здесь:

```
Σ_w δ_w= 1
```

это. δw:

- до получения результата,
- в соответствии с временным весом

необходимо определить. Накопительная оценка:

- текущая оценка,
- не заменяет историческую запись инцидентов

Ни одна запись не может быть тихо заменена.

50. ТРИ ОЦЕНКИ ДЛЯ НЕРАЗРЕШЕННЫХ ТРЕБОВАНИЙ

Неразрешенные требования могут искусственно увеличивать или уменьшать точечную оценку материальных единиц. Поэтому предлагается три результата.

50.1. Консервативная нижняя граница

Неразрешенные требования считаются нулевыми. Медленнее

50.2. Точечная оценка только для разрешенных

Используются только разрешенные требования. Spoint

50.3. Оптимистичная верхняя граница

Неразрешенные требования принимаются с полной поддержкой. Supper Следовательно:

```
S_lower≤ S_point≤ S_upper
```

должно быть так. Этот диапазон:

это диапазон справочной или эпистемической неопределенности.

Это не то же самое, что доверительный интервал выборки.

51. НЕОПРЕДЕЛЕННОСТЬ ВЫБОРКИ

Поскольку оценки по популяции получаются из выборки:

- распределение откликов,
- оценка компонента,
- составная оценка

должны иметь доверительный интервал. Метод в сложных выборках:

- слои,
- наборы,
- их веса,
- сочетанные пользователи,
- волны

должны быть защищены. Возможные методы:

- Оценка дисперсии на основе дизайна
- Реплицированные веса
- Слоистый бутстрэп
- Кластерный бутстрэп
- Соответствующая модель для редких событий
- Оценка опроса с помощью модели

Используемый метод должен быть указан в публичной документации.

52. ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ СОСТАВНОГО БАЛЛА

Доверительный интервал составного балла нельзя получить простым сложением интервалов компонентов. Компоненты:

- те же пользователи,
- те же утверждения,
- те же языки,
- те же волны

зависит от этой причины. Предпочтительный подход: создавать повторные выборки в соответствии с дизайном выборки. В каждой повторной выборке пересчитывайте все показатели отклика и компоненты. Воссоздавайте составной балл. Получайте доверительный интервал из распределения. Этот процесс должен заново запускать всю архитектуру оценки.

53. ЭФФЕКТИВНЫЙ РАЗМЕР ВЫБОРКИ

Сырые наблюдения: n и эффективная выборка:

```
n_eff= (Σ_i w_i)^2/ Σ_i w_i^2 [K08]
```

должны публиковаться вместе. Сырые наблюдения 1,000 не обязательно означают, что 1,000 равны единицам информации. Оценки компонентов и подгрупп должны иметь собственные эффективные размеры выборки.

54. ОБНАРУЖЕНА КРИТИЧЕСКАЯ ОШИБКА НУЛЕВОГО ЗНАЧЕНИЯ

Отсутствие Подтвержденных Критических событий в выборке:

Наблюдаемое Подтвержденное Критическое = 0

должно быть написано. Следующее выражение не должно использоваться:

> *Критический риск равен нулю.*

При допущении независимых и одинаковых наблюдений быстрый приближённый верхний предел:

```
p_upper,95≈ 3/n [K09]
```

можно оценить с известным подходом. В сложном проектировании GEO-1000:

- Вместо исходного n следует использовать метод, соответствующий проектированию,
- при необходимости эффективную выборку,
- точный или основанный на бутстрепе односторонний предел

должен быть использован. Пример:

```
neff=1,000
```

и приближённый верхний предел при нулевых событиях:

```
3/1,000 = 0.003 = 0.3%
```

может быть. Корректное публичное заявление:

> *В 1,000 действительных наблюдений не было зафиксировано ни одного подтвержденного критического события; в используемом приближенном методе односторонний верхний предел 95 составляет примерно 0.3 процентов.*

55. НЕОПРЕДЕЛЕННОСТЬ ПРИ ОДНОМ КРИТИЧЕСКОМ СОБЫТИИ

Когда наблюдается одно критическое событие: точечная оценка может быть низкой, доверительный интервал может быть широким. Корректный результат:

- событие произошло,
- оцененная распространенность,
- широкая неопределенность,
- текущее состояние

должны быть показаны вместе. Следующие два предложения также могут быть неверными:

- «Только одно событие незначительно.»
- «Событие доказывает, что вся система потерпела неудачу 100%.»

56. ЧУВСТВИТЕЛЬНОСТЬ ОЦЕНКИ

Также могут быть рассчитаны следующие альтернативы:

- Сырой несбалансированный балл
- Баллы с взвешиванием по основной популяции
- Балл с невыровненным весом
- Балл с усечённым весом
- Балл только для решённых случаев
- Консервативный балл
- Балл с равным произведением
- Балл с весом экспозиции
- Оценка Естественный охват
- Оценка Общая опора сравнения
- Чистый балл
- Естественный балл

Если существует существенная разница:

ПРЕДУПРЕЖДЕНИЕ О ЧУВСТВИТЕЛЬНОСТИ ОЦЕНКИ

должно быть предоставлено.

57. НЕПРАВИЛЬНАЯ ТОЧНОСТЬ

Следующее представление может быть чрезмерно точным для большинства результатов GEO-1000: 963,742/ 1000 Точечная оценка:

- целое число,
- десятичное число при необходимости

должно быть опубликовано. Доверительный интервал должен указывать предел числовой точности.

58. ОТСУТСТВУЮЩИЙ КОМПОНЕНТ

Если один из десяти компонентов не может быть вычислен:

- его вес не может быть распределен на другие компоненты.
- отсутствующее поле не может быть засчитано как 100 баллов,

полный NOMOS результат не может быть опубликован. Правильный статус:

ЧАСТИЧНЫЙ NOMOS ПРОФИЛЬ

КОМПОНЕНТ НЕ ОЦЕНИВАЕТСЯ

ПРОВИЗОРНЫЙ РЕЗУЛЬТАТ

может применяться. Пример: если доступна не вся панель для USI, можно представить только частичный профиль с восемью или девятью компонентами.

```
USI=100 не может быть принято как данное.
```

59. СТАТУС ПУБЛИКАЦИИ РЕЗУЛЬТАТА

NSQ-0— НЕ ВЫЧИСЛЕНО

Необходимые данные недоступны.

NSQ-1— ПРОВИЗОРНЫЙ

Существует одна волна, ограниченное урегулирование или отсутствующие компоненты.

NSQ-2— ЧАСТИЧНЫЙ ПРОФИЛЬ

Некоторые компоненты были надежно рассчитаны. Полное сравнение по точкам 1,000 невозможно.

NSQ-3— ДЕЙСТВИТЕЛЬНЫЙ БАЛЛ ВОЛНЫ

Все ключевые компоненты и предварительные условия качества выполнены.

NSQ-4— ПОВТОРНЫЙ БАЛЛ

Баллы были повторены хотя бы в одной независимой или подтверждающей волне.

NSQ-5— ВНЕШНЕ АУДИРОВАННЫЙ БАЛЛ

Расчет и цепочка доказательств были воспроизведены независимой командой. Кандидатура 950+ по крайней мере:

NSQ-4

должен нацеливаться на этот уровень.

60. ВЕРСИОНИРОВАНИЕ СЧЁТА

Каждая версия счёта должна включать следующее:

- Версия кода вычислений
- Формулы компонентов
- Сопоставления значений измерений
- Весовые коэффициенты семейств промптов
- Коэффициенты продукта
- Весовые коэффициенты образцов
- Версия Truth Pack
- Версия арбитра
- ID волны
- Хэш целостности

Если формула изменяется: старый счёт не может быть переписан, создаётся новая версия счёта.

61. НОВАЯ ФОРМУЛА СЧЁТА ДЛЯ ТЕХ ЖЕ ДАННЫХ

Методика оценки может эволюционировать. Те же исходные данные:

- Метод оценки 0.9

- Метод оценки 1.0

можно пересчитать с. Корректная запись:

Данные	Метод оценки	Результат
Фиксированный набор наблюдений W1	0.9	944
Фиксированный набор наблюдений W1	1.0	951

Два результата являются отдельными версиями. Новый метод не может удалить старый метод из прошлого.

62. ЗАЩИТА ОТ МАНИПУЛЯЦИЙ ОЦЕНКОЙ

Архитектура оценки NOMOS должна иметь явную защиту против следующих методов:

- Выбор только продуктов ИИ с высокой оценкой
- Исключение языков или стран с низкой оценкой из области
- Преобразование Условное соответствие в Полное соответствие
- Учёт невыполненных как положительных
- Молчаливое удаление Не подлежит оценке из знаменателя
- Изменение веса семейства промптов после получения результата
- Уменьшение коэффициента большого продукта в зависимости от оценки
- Сокрытие продукта с критическим событием в среднем по экосистеме
- Публикация только чистой оценки
- Выбор лучшей волны в качестве текущей оценки
- Сокрытие разницы между несвешенной и взвешенной оценками
- Распределение веса отсутствующих компонентов на сильные компоненты
- Не предоставление интервала доверия
- Прохождение порога 950 на основе точечной оценки и скрытие неопределенности

63. ПУБЛИЧНАЯ КАРТА РЕЗУЛЬТАТОВ NOMOS

Публичная карта результатов кандидата должна содержать следующий порядок:

A. Идентичность

Аудируемый объект ИИ-продукт Естественный охват/ Общая опора сравнения Контролируемый / Естественный набор промптов Wave Truth Pack версия Метод оценки

B. Статус шлюза

Полное соответствие Кандидат Условный Существенное несоответствие Критическая блокировка Не подлежит оценке

C. Распределение ответов GEO-1000

RP-1–RP-8 эквиваленты пользователей Строгий проход Допустимый проход Критический Основной Не решено Нет пригодного отклика Не подлежит оценке

D. Профиль из десяти компонентов

ENT

FAC

SBI

TLA

ECI

EPI

TCR

LGF

USI

STR

E. Составной результат

NOMOS/ 1,000

Интервал доверия выборки Опорные нижняя/верхняя границы Эффективная выборка

F. Риски и события

Наблюдаемые подтвержденные критические события; критическая верхняя граница; основные выводы; открытые инциденты; и статус исправления.

G. Объем и качество

SQ

PSI

NCL

TPR

AQ

NSQ

Охват по странам/языкам Уровень сравнения

64. РАСПРЕДЕЛЕНИЕ СИНТЕТИЧЕСКИХ ОТВЕТОВ

ДЕМОНСТРАЦИЯ СИНТЕТИЧЕСКОЙ МЕТОДОЛОГИИ / Следующие результаты не являются реальными продуктами ИИ или показателями компании. Для Synthetic Orion AI:

Статус	Эквивалент пользователя
RP-1 Полное соответствие	730
RP-2 Сдано с замечаниями	170
RP-3 Условное соответствие	70
RP-4 Существенное несоответствие	20
RP-5 Критическое несоответствие	2
RP-6 Не решено	5
RP-7 Нет пригодного ответа	3
RP-8 Не подлежит оценке	0
Итого	1,000

Строгая сдача:

```
730+170=900
```

Допустимая сдача:

```
900+70=970
```

Критично: 2/1,000 Важно: 20/1,000 Допустимый уровень сдачи для этого продукта составляет 97 процентов. Однако есть явные критические и важные замечания. Корректный статус:

Критическая блокировка

Некорректный статус:

97 процентов успешно

65. ПРОФИЛЬ СИНТЕТИЧЕСКОГО СОЕДИНЕНИЯ

Компонент	Оценка
ENT	96
FAC	93
SBI	89
TLA	95
ECI	88
EPI	91
TCR	96
LGF	82
USI	87
STR	90
NOMOS	907

Правильное отображение для публики:

Критическая блокировка— 907/1,000

Эта система:

- в среднем сильные области,
- слабый язык/справедливость страны,
- вопросы материальных доказательств,
- открытое критическое событие

несет.

66. СИНТЕТИЧЕСКИЙ РЕЗУЛЬТАТ ПОСЛЕ КОРРЕКЦИИ

В новой волне:

Статус	Эквивалент пользователя
RP-1	830
RP-2	128
RP-3	30
RP-4	0
RP-5	0
RP-6	5
RP-7	7
RP-8	0
Итого	1,000

Строгий проход: Компоненты 958:

Компонент	Оценка
ENT	99
FAC	97
SBI	96
TLA	95
ECI	96
EPI	98
TCR	97
LGF	94
USI	93
STR	96
NOMOS	961

Правильная запись: ТЕКУЩАЯ: NOMOS 950+ КАНДИДАТ НА ПРОИЗВОДИТЕЛЬНОСТЬ — 961/1,000/ Историческая: Ранее подтвержденный критический инцидент устранен / Наблюдаемый критический при текущем повторном тестировании: 0/ Верхняя граница критической частоты: также показано Старый критический инцидент не удален.

67. СИНТЕТИЧЕСКИЙ AI ПРОДУКТОВЫЙ КОЭФФИЦИЕНТ СЛУЧАЯ

Четыре синтетических AI продукта:

Продукт	NOMOS	Коэффициент равенства	Коэффициент воздействия
A1	960	0.25	0.55
A2	940	0.25	0.25
A3	900	0.25	0.15
A4	980	0.25	0.05

Счет равного продукта:

```
0.25(960+940+900+980)=945
```

Счет с учетом веса воздействия:

```
0.55(960)+0.25(940)+0.15(900)+0.05(980)=947
```

Два результата близки. Однако они могут сильно отличаться при другой распределении. Правильная интерпретация:

- Счет экосистемы равной системы: 945
- Счет экосистемы с учетом веса воздействия: 947
- Записи по продукту: отдельные данные о вратарях и инцидентах

Неправильный комментарий: «Оценка экосистемы NOMOS определенно 947.»

68. СИНТЕТИЧЕСКИЙ СЛУЧАЙ КРИТИЧЕСКОГО КОЭФФИЦИЕНТА

Пусть оценка продукта A4 равна 980. Однако пусть в A4 произойдет Подтвержденное Критическое событие. Коэффициент воздействия всего лишь 0.05. Композитная оценка экосистемы может оставаться высокой. Правильная публичная карта: Композит с учетом воздействия: 947/ Любой Подтвержденный Критический продукт: ДА / Доля критического воздействия: вычисленное соотношение / Покрытие A4: Критическая блокировка. Малый коэффициент A4 не может стереть Критическое событие.

69. РЕШЕНИЕ О СТАТУСЕ ПУБЛИКАЦИИ ОЦЕНКИ

Порядок выбора кандидата:

```
PublicationStatus= Не подлежит оценке(отсутствует базовое измерение); Критическая
блокировка(открыто Подтверждено Критическое); Существенное несоответствие(открыто Подтверждено
Основное); PROVISIONAL(отсутствует компонент, повторение или качество); CONDITIONAL(Умеренное или
порог неразрешенного); 950+ CANDIDATE(все кандидаты на требования); в противном случае ДОПУСТИМО
```

Диапазон оценок показан после этого решения о статусе.

70. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH16-N01

Результат NOMOS не может быть опубликован как отдельная чистая составная оценка.

CH16-N02

Статус шлюза должен быть показан перед составной оценкой.

CH16-N03

Распределение ответов GEO-1000 является обязательной частью основного публичного результата.

CH16-N04

Сумма эквивалентов пользователей RP-1 и RP-8 должна быть 1,000.

CH16-N05

Строгий проход, Допустимый проход, Основной, Критический, Не решено, Нет используемого ответа и Не подлежит оценке должны отображаться отдельно.

CH16-N06

Невыясненные наблюдения и наблюдения Не подлежит оценке не могут быть тихо добавлены в категории «проход» или «неудача».

CH16-N07

Наблюдения Не подлежит оценке не могут интерпретироваться как сбой в работе продукта.

CH16-N08

Наблюдения Не подлежит оценке не могут быть тихо удалены из знаменателя; их следует показывать в оценочном охвате.

CH16-N09

Вес наблюдения должен быть независимым от результата ответа ИИ.

CH16-N10

Сырые данные о количестве участников и эквивалент пользователя с учетом веса населения должны сообщаться отдельно.

CH16-N11

Статус прохождения отклика должен храниться отдельно от коэффициента поддержки атомарного утверждения.

CH16-N12

Критические и основные ворота не могут компенсироваться никаким составным баллом.

CH16-N13

Вес атомарной центральности должен быть определен до просмотра результатов.

CH16-N14

Повторение того же корневого утверждения не может искусственно увеличивать или уменьшать оценку.

CH16-N15

Вес центральности утверждения не может заменять критические и основные ворота.

CH16-N16

Значения измерения атомарной точности должны иметь версионирование и быть общедоступными.

CH16-N17

Если карта атомарной точности меняется, должна быть создана новая версия метода оценки.

CH16-N18

Атомы Пробел в эталонных данных не могут быть тихо добавлены к оценочному баллу как ноль или один.

CH16-N19

Полнота обязательного элемента ответа должна публиковаться как отдельный показатель.

CH16-N20

Сбой основного обязательного элемента не может быть компенсирован высоким средним показателем REC.

CH16-N21

Профиль NOMOS с десятью компонентами должен быть виден перед составным индексом.

CH16-N22

ENT, FAC, SBI, TLA, ECI, EPI, TCR, LGF, USI и STR должны иметь отдельные оценки.

CH16-N23

Одна компонента не может молчаливо взять на себя недостающий вес другой компоненты.

CH16-N24

Компонента, которую нельзя рассчитать, не может считаться 100.

CH16-N25

Полный составной индекс NOMOS должен публиковаться только если все десять компонентов были надежно рассчитаны.

CH16-N26

Если используется равное взвешивание компонентов в качестве канонического метода, это должно быть явно указано.

CH16-N27

Альтернативные веса для конкретной области должны иметь отдельное название и версию оценки.

CH16-N28

Оценка, специфичная для конкретной области, не может использоваться вместо канонического индекса без объяснения.

CH16-N29

LGF не может опираться только на среднее значение по населению; она должна включать базу производительности и различия.

CH16-N30

Одно наблюдение Наблюдателя страны не может быть преобразовано в оценку справедливости страны.

CH16-N31

В оценке справедливости должны участвовать только языковые и географические группы с достаточным размером выборки.

CH16-N32

Языки и страны, которые не были протестированы, нельзя игнорировать при получении высоко справедливого результата.

CH16-N33

USI должен оценивать результаты как контролируемой, так и естественной панели вместе с различиями в интерфейсе материала.

CH16-N34

В исследовании с только контролируемой или только естественной панелью USI не может считаться полностью рассчитанным.

CH16-N35

Баллы контролируемой и естественной панели по умолчанию должны храниться отдельно.

CH16-N36

Если баллы контролируемой и естественной панели объединяются, коэффициенты должны быть зафиксированы до получения результатов.

CH16-N37

STR не может быть представлен как полный показатель стабильности в одной волне.

CH16-N38

Изменения продукта между волнами или в Truth Pack должны явно моделироваться при расчете стабильности.

CH16-N39

Соединение NOMOS не может изменять статус критического или основного шлюза.

CH16-N40

Оценка 950 сама по себе не может установить кандидатуру NOMOS 950+.

CH16-N41

Условия кандидата NOMOS 950+ должны совместно соответствовать требованиям по оценке, строгому прохождению, шлюзу, минимальному компоненту, неразрешенным вопросам, качеству и повторяемости.

CH16-N42

Пороги 950+ не могут быть представлены в качестве окончательного научного стандарта до пилотного и независимого обзора.

CH16-N43

Идентификация оценки должна включать продукт ИИ, рамку, панель, набор промптов, волну, Truth Pack и версию оценки.

CH16-N44

Предложение «Оценка NOMOS сущности составляет X» не может использоваться без идентичности области.

CH16-N45

Результат Core GEO-1000 не может быть представлен как полный балл NOMOS.

CH16-N46

Баллы Естественный охват и Общая опора сравнения должны быть представлены отдельно.

CH16-N47

Баллы продуктов для разных групп Естественный охват не могут быть напрямую ранжированы без пояснений.

CH16-N48

Результат Общая опора сравнения не может быть представлен как полный глобальный доступный балл продукта.

CH16-N49

Коэффициенты продукта ИИ должны определяться независимо от результатов.

CH16-N50

Коэффициент продукта ИИ не может зависеть от балла продукта, репутации бренда или коммерческих спонсорских отношений.

CH16-N51

Равные баллы продукта и экосистемы с учетом воздействия должны иметь отдельные наименования.

CH16-N52

Если данные об активных пользователях или о воздействии не могут быть проверены, произвольные коэффициенты не могут быть сгенерированы.

CH16-N53

Данные о воздействии должны быть прозрачными на уровне аккаунта, пользователя, запроса или сессии.

CH16-N54

Коэффициент продукта с низкой оценкой не может быть уменьшен после получения результата.

CH16-N55

Составная экосистема не может удалить Критический или Важный шлюз на уровне продукта.

CH16-N56

Коэффициент AnyCritical и критический коэффициент с учетом воздействия должны отображаться отдельно.

CH16-N57

Коэффициенты семейства промптов должны определяться до получения результатов.

CH16-N58

Вес семейства промптов с низкой оценкой не может быть уменьшен позже.

CH16-N59

Отдельный результат каждой группы промптов должен оставаться видимым вместе с составным баллом промпта.

CH16-N60

Текущий волновой балл должен храниться отдельно от скользящих или исторических баллов.

CH16-N61

Прошлая волна с наивысшим баллом не может быть выбрана в качестве текущего балла.

CH16-N62

Для массы нерешённых требований могут быть получены консервативные, только решённые и оптимистичные результаты.

CH16-N63

Определённость ссылки и доверительный интервал выборки не могут быть представлены как одно и то же.

CH16-N64

Сложный доверительный интервал выборки должен сохранять стратификацию, кластеризацию, взвешивание и структуру сопоставленных пользователей.

CH16-N65

Интервал доверия составного балла должен быть рассчитан путем повторного запуска всего процесса оценки.

CH16-N66

Следует показывать необработанные и эффективные размеры выборки вместе.

CH16-N67

Критическое событие, наблюдаемое как ноль, не может быть выражено как нулевой критический риск.

CH16-N68

В случае нулевых событий следует опубликовать односторонний верхний предел или эквивалентную неопределённость редкого события.

CH16-N69

Подход 3/n не может использоваться автоматически в качестве окончательного метода в сложном проектировании; он может быть только приблизительным управлением.

CH16-N70

В критическом событии следует одновременно показывать наличие события и неопределенность его частоты.

CH16-N71

Анализы чувствительности оценки должны показывать результаты по невзвешенным, взвешенным, усечённым, нерешённым и альтернативным коэффициентам в той мере, в какой это актуально.

CH16-N72

Чувствительность материальной оценки должна создавать предупреждение на публичной карточке.

CH16-N73

Фальшивое десятичное значение оценки не может публиковаться вообще.

CH16-N74

Вес отсутствующего компонента не может быть распределён между другими компонентами.

CH16-N75

Статус качества публикации Puan должен быть виден в публичной записи.

CH16-N76

Кандидатура 950+ должна стремиться как минимум к уровню воспроизведённой оценки.

CH16-N77

Когда формула оценки меняется, прошлые оценки не могут быть переписаны без уведомления.

CH16-N78

Если к одному и тому же набору данных применяются разные версии метода оценки, все результаты должны сохраняться в отдельных версиях.

CH16-N79

Код расчета оценки, коэффициенты и записи целостности должны публиковаться в формате, доступном для машинного чтения.

CH16-N80

Каждая оценка, коэффициент, доверительный интервал и решение о публикации должны иметь ответственного человека или организацию.

71. ФОРМЫ ОШИБОК

CH16-F01— ОТЧЕТ С ОДНИМ ЧИСЛОМ

Распределение, порог и профиль компонентов сохраняются.

CH16-F02— СКРЫТИЕ ПОРГА ПОД ОЦЕНКОЙ

Критическая блокировка становится небольшой сноской.

CH16-F03— УЧЕТ УСЛОВНЫХ КАК Полное соответствие

Умеренные проблемы становятся незаметными.

CH16-F04— ПОДСЧЕТ НЕРАЗРЕШЕННЫХ КАЖДЫМ ПРОХОДОМ

Неизвестные результаты повышают оценку.

CH16-F05— ПОДСЧЕТ НЕРАЗРЕШЕННЫХ КАК СБОЙ

Отсутствие ссылки снижает продукт.

CH16-F06— УДАЛЕНИЕ Не подлежит оценке ИЗ ЗНАМЕНАТЕЛЯ

Дефекты измерения становятся незаметными.

CH16-F07— ПОДСЧЕТ Не подлежит оценке КАК ДЕФЕКТА ПРОДУКТА

Инспекционный дефект назначается AI.

CH16-F08— ПОДСЧЕТ 1,000 ЛЮДЕЙ С РАВНЫМ ВЕСОМ

Игнорируются передвыборка и различия в популяции.

CH16-F09— ВЕС В СООТВЕТСТВИИ С РЕЗУЛЬТАТОМ

Правильные ответы получают больший вес.

CH16-F10— ПОДСЧЕТ Доля подтверждённых утверждений КАК ПРОХОДНОЕ РЕШЕНИЕ

Критический элемент теряется в среднем значении.

CH16-F11— УЧИТЫВАТЬ ОСНОВНЫЕ И ВСПОМОГАТЕЛЬНЫЕ ОДИНАКОВО

Вспомогательное утверждение получает такой же вес, как и основной объект.

CH16-F12— УВЕЛИЧИВАТЬ ОЦЕНКУ ПОВТОРЕНИЕМ

То же правильное утверждение записывается десять раз для увеличения балла.

CH16-F13— СНИЖАТЬ ОЦЕНКУ ПОВТОРЕНИЕМ

Тот же неправильный корень приводит к десяти отдельным наказаниям.

CH16-F14— ХРАНЕНИЕ ЗНАЧЕНИЙ ИЗМЕРЕНИЙ

Оценка не может быть воспроизведена.

CH16-F15— УЧЕТ УТВЕРЖДЕНИЯ, ПРИШЕДШЕГО СНИЗУ, КАК ФАКТА ДЛЯ F-4

Точная передача самооценки превращается в точность самооценки.

CH16-F16— ПЛАВЛЕНИЕ НЕОБХОДИМЫХ ЭЛЕМЕНТОВ В АТОМНОЕ СРЕДНЕЕ

Даже если основная задача отсутствует, ответ получает высокий балл.

CH16-F17— ВЫБОР СЛАДКИХ КОМПОНЕНТОВ

Публикуются только три сильные компоненты.

CH16-F18— УЧЁТ ОТСУТСТВУЮЩЕГО КОМПОНЕНТА КАК 100

Полный балл искусственно увеличен.

CH16-F19— РАСПРЕДЕЛЕНИЕ ОТСУТСТВУЮЩЕГО ВЕСА

Нерассчитанный показатель справедливости добавлен к фактическому баллу.

CH16-F20— УЧЁТ ЗАКОНА ПРИРОДЫ КАК РАВНОГО ВЕСА

Чувствительность и внешняя проверка не выполняются.

CH16-F21— УЧЁТ ОЦЕНКИ ДОМЕНА КАК КАНОНИЧЕСКОЙ

Особый вес представлен как базовый балл NOMOS.

CH16-F22— УЧЁТ СРЕДНЕГО ЯЗЫКА КАК ЯЗЫКОВОЙ СПРАВЕДЛИВОСТИ

Низкий языковой порог скрыт.

CH16-F23— УЧЁТ РАВНОСТИ ЯЗЫКОВ КАК КАЧЕСТВА

Низкие оценки на всех языках дают высокую справедливость.

CH16-F24— УЧЁТ НЕПРОТЕСТИРОВАННЫХ ЯЗЫКОВ КАК ПОКРЫТИЯ

Фактор покрытия завышен.

CH16-F25— ОЦЕНКА СТРАНЫ ПО ОЦЕНЩИКУ

Один пользователь вносит вклад в учёт справедливости.

CH16-F26— ПУБЛИКАЦИЯ ТОЛЬКО ЧИСТОЙ ОЦЕНКИ

Естественный пользовательский опыт сохраняется.

CH16-F27— УЧИТЫВАТЬ ТОЛЬКО ЕСТЕСТВЕННУЮ ОЦЕНКУ КАК СТАНДАРТ ПРОДУКТА

Персонализация становится базовым поведением.

CH16-F28— ТИХО СРЕДНЕ ЧИСТО И ЕСТЕСТВЕННО

Разные объекты предсказания смешиваются.

CH16-F29— 100 СТАБИЛЬНОСТЬ В ОДНОЙ ВОЛНЕ

Стабильность считается полной без повторения.

CH16-F30— СЧИТАЙ СТАРУЮ ХОРОШУЮ ВОЛНУ КАК ТЕКУЩУЮ

Текущее ослабление скрыто.

CH16-F31— АВТОМАТИЧЕСКИ СЧИТАЙ 950 КАК СООТВЕТСТВИЕ

Ворота и пол компонента игнорируются.

CH16-F32— Критическая блокировка— СЧИТАЙ 970 КАК 950+

Счет опережает ворота.

CH16-F33— СКРЫТЬ ИДЕНТИЧНОСТЬ ОЦЕНКИ

Неизвестно, какой продукт и панель измеряются.

CH16-F34— СЧИТАТЬ ОСНОВНОЙ БАЛЛ ПОЛНЫМ NOMOS

Одна промпт представляет весь стандарт.

CH16-F35— РАНЖИРОВАТЬ ОЦЕНКИ Естественный охват

Разные популяции считаются одинаковыми.

CH16-F36— СЧИТАТЬ ПОЛНЫЙ ДОСТУП К Общая опора сравнения

Узкая общая популяция становится глобальным результатом.

CH16-F37— ПРЕДОСТАВИТЬ КОЭФФИЦИЕНТ МОДЕЛИ В СООТВЕТСТВИИ С ОЦЕНКОЙ

Высокий балл получает больше влияния.

CH16-F38— ДЕЛАЕМ РЕПУТАЦИЮ БРЕНДА ФАКТОРОМ

Популярность продукта учитывается без прогнозирования.

CH16-F39— УЧЁТ КОЛИЧЕСТВА АКТИВНЫХ ПОЛЬЗОВАТЕЛЕЙ ПО НОМЕРАМ СЧЁТОВ

Коэффициент экспозиции увеличивается.

CH16-F40— УДАЛЕНИЕ ПРОДУКТОВ С НИЗКИМ РЕЙТИНГОМ

Экосистема состоит только из успешных продуктов.

CH16-F41— УЧЁТ ОДИНАКОВОГО СЧЕТА ЭКСПОЗИЦИИ ДЛЯ ПРОДУКТА

Средний продукт представлен как средний человек.

CH16-F42— УЧЁТ СЧЕТА ЭКСПОЗИЦИИ КАК КАЧЕСТВА ПРОДУКТА

Большой продукт автоматически становится более важным показателем качества.

CH16-F43— УДАЛИТЬ КРИТИЧНОЕ В СРЕДНЕМ ЭКОСИСТЕМЫ

Тяжелое событие продукта с малым коэффициентом исчезает.

CH16-F44— ИЗМЕНИТЬ КОЭФФИЦИЕНТ промпта ПОСЛЕ РЕЗУЛЬТАТА

Семья жестких промптов сокращается.

CH16-F45— РАССМАТРИВАТЬ СЕМЬЮ С НИЗКИМ БАЛЛОМ КАК ОПЦИОНАЛЬНУЮ

Пограничный или доказательный результат исключается из области.

CH16-F46— СОХРАНЯТЬ ТЕКУЩЕЕ СО СКОЛЬЗЯЩИМ БАЛЛОМ

Старые высокие волны покрывают текущее падение.

CH16-F47— УДАЛИТЬ НЕРАЗРЕШЕННЫЕ ИЗ ЗНАМЕНАТЕЛЯ

Точечная оценка искусственно повышается.

CH16-F48— УЧИТЫВАЙТЕ НЕОПРЕДЕЛЕННОСТЬ ССЫЛКИ ПРИ ВЫБОРОЧНОМ ИНТЕРВАЛЕ ДОВЕРИЯ

Смешиваются две разные неопределенности.

CH16-F49— ПРОСТОЙ СЛУЧАЙНЫЙ ВЫБОР ИНТЕРВАЛА ДОВЕРИЯ

Стратификация, вес и кластеры игнорируются.

CH16-F50— СУММИРОВАНИЕ КОМПОНЕНТНЫХ ИНТЕРВАЛОВ ДОВЕРИЯ

Зависимость игнорируется.

CH16-F51— УТВЕРЖДЕНИЕ ЧУВСТВИТЕЛЬНОСТИ С НЕОБРАБОТАННЫМ N

Эффективная выборка скрыта.

CH16-F52— ПРИСВАИВАНИЕ НУЛЯМ СОБЫТИЙ НУЛЕВОЙ РИСК

Неопределенность редкой критической ошибки скрыта.

CH16-F53— ПРИМЕНЕНИЕ 3/N К КАЖДОМУ СЛУЧАЮ

Сложная структура выборки и весов игнорируется.

CH16-F54— УЧЁТ ОДНОГО КРИТИЧЕСКОГО КАК 100 В СИСТЕМЕ РИСКА В ПРОЦЕНТАХ

Общность чрезмерно обобщена.

CH16-F55— СКРЫТИЕ ОДНОГО КРИТИЧЕСКОГО, ПОТОМУ ЧТО СТАВКА МАЛА

Существование события исчезает.

CH16-F56— СКРЫТИЕ АНАЛИЗА ЧУВСТВИТЕЛЬНОСТИ

Не видно, что оценка изменяется при альтернативных разумных методах.

CH16-F57— ЛОЖНАЯ ДЕСЯТИЧНАЯ ЧУВСТВИТЕЛЬНОСТЬ

Публикуется результат вроде 963,742.

CH16-F58— УЧЁТ ЧАСТИЧНОГО ПРОФИЛЯ КАК ПОЛНОГО БАЛЛА

Сравнение проводится с отсутствующим компонентом.

CH16-F59— ПУБЛИКАЦИЯ ПРОВИЗОРНОГО БАЛЛА КАК ОКОНЧАТЕЛЬНОГО

Одна волна и один арбитр становятся полной нормой.

CH16-F60— ПЕРЕПИСЫВАНИЕ ИСТОРИИ ПРИ ИЗМЕНЕНИИ ФОРМУЛЫ

Старый балл исчезает.

CH16-F61— СОХРАНЯТЬ КОД ВЫЧИСЛЕНИЙ ЗАКРЫТЫМ

Независимое воспроизведение предотвращается.

CH16-F62— СЧИТАТЬ ВЫСОКИЙ БАЛЛ КАЧЕСТВОМ АУДИТА

Принято неисправное измерение, так как оно дало высокий результат.

CH16-F63— СЧИТАТЬ НИЗКИЙ БАЛЛ ДЕФЕКТОМ МЕТОДА

Низкий результат пересчитывается из-за коммерческих неудобств.

CH16-F64— МОНОПОЛИЯ ОДНОЙ КОМПАНИИ ПО ОЦЕНКЕ

Независимым учреждениям запрещается применять ту же формулу.

72. ПРОЦЕДУРА АУДИТА

Шаг 1— Создать идентификатор оценки

Продукт AI сущности Frame Panel Prompt, набор Wave Truth Pack методом оценки зафиксирован.

Шаг 2— Зафиксировать основной набор наблюдений анализа

Резервные, отозванные, не подлежащие оценке и основные записи наблюдений разделены.

Шаг 3— Проверить веса населения

Дизайн, неответ и калибровочные веса проверены.

Шаг 4— Рассчитать распределение RP

Пользовательские эквиваленты с RP-1 до RP-8 созданы.

Шаг 5— Рассчитать строгий и приемлемый проход

SP1 000 и AP1 000 получены.

Шаг 6— Рассчитать плотности критических, основных, нерешённых и неоценимых

Шлюз и распространённость сохраняются отдельно.

Шаг 7— Применить центральность утверждений и нормализацию корневых кластеров

утверждения проверяются снова.

Шаг 8— Рассчитать метрики атомного размера

Производятся результаты по сущности, факту, объему, времени, атрибуции, модальности и ссылке.

Шаг 9— Вычислить требуемую полноту элемента

Он фиксируется вместе с семейством шлюзов промптов.

Шаг 10— Вычислить десять компонентов

Применяются все формулы компонентов от ENT до STR.

Шаг 11— Проверить наличие отсутствующих компонентов

Определяется, можно ли вычислить полный балл.

Шаг 12— Вычислить составной индекс NOMOS

Применяется только к полным и действительным профилям.

Шаг 13— Назначить статус шлюза

Определяет, является ли статус Критическим, Основным, Условным или Полным кандидатом.

Шаг 14— Разделение результатов Естественный охват и Общая опора сравнения

Две семейства баллов хранятся отдельно.

Шаг 15— Фиксация коэффициентов продукта AI

Коэффициенты равенства и воздействия сохраняются вместе с источником данных.

Шаг 16— Применение коэффициентов семейства Prompt

Весовые коэффициенты используются независимо от результата.

Шаг 17— Разделение результатов панели и волны

Clean, Natural, Current и Rolling баллы разделяются.

Шаг 18— Расчет показателей справедливости и минимального уровня

Создаются результаты по языку, стране и охвату.

Шаг 19— Расчет неразрешенных результатов нижней–средней–верхней точек

Справочная неопределенность делается видимой.

Шаг 20— Расчет доверительных интервалов выборки

Применяется метод повторной выборки, подходящий для структуры выборки.

Шаг 21— Расчет эффективного объема выборки

Отображается на общей и компонентной основе.

Шаг 22— Расчет порога редкого критического события

Используется односторонний метод для нуля или нескольких событий.

Шаг 23— Генерация анализа чувствительности

Сравниваются вес, коэффициент и нерешённые варианты.

Шаг 24— Назначение статуса качества публикации

Принимается решение между NSQ-0 и NSQ-5.

Шаг 25— Применение NOMOS 950+ кандидатских дверей

Проверяются все условия, кроме оценки.

Шаг 26— Создание карты публичных результатов

Ворота, распределение, профиль, индекс, неопределённость и область публикуются соответственно по порядку.

Шаг 27— Фиксация вычислительного кода и манифеста

Версия кода, входные данные, хеш и результаты связаны.

Шаг 28— Подготовка пакета для независимого воспроизведения

Файлы вычислений, очищенные от персональных данных, предоставляются для независимой проверки.

73. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Идентификатор оценки Сущность запись ИИ Система Регистрация Естественный охват и Общая опора сравнения записи Статус панели Набор респондентов и семейства Версия пакета

истины Идентификатор волны Книга требований Решения по ответам RP-1–RP-8 распределение Основной анализ Набор наблюдений Дизайнерские веса Веса неответов Калибровочные веса Обрезка весов записи Эффективная выборка Строгий проход Приемлемый проход Критическая ставка Основная ставка Нерешённая ставка Нет Используемых ставок ответа Не подлежит оценке ставка Веса централизации требований Основной кластер записи Значения атомных измерений Атомная точность Требуемое завершение элементов

ENT

FAC

SBI

TLA

ECI

EPI

TCR

LGF

USI

STR

Компонентные подвходы; минимальные значения для языка и страны; коэффициент покрытия; контролируемые и естественные оценки; результаты между волнами; составной индекс NOMOS; альтернативные оценки по конкретной области; коэффициенты ИИ-продукта и их источники данных; коэффициенты семейства промптов; результат для одинакового продукта; результат с учетом взвешивания по экспозиции; статус AnyCritical; доля критической экспозиции; неразрешенные результаты нижняя–точка–верхняя; доверительный интервал выборки; верхняя критическая граница; анализ чувствительности; качество публикации оценки; запись кандидата-шлюза 950+; версия метода оценки; код расчета; хэш целостности кода; публичная таблица оценок; журнал изменений; запись независимого воспроизведения; и ответственное лицо или организация.

74. ЧЕК-ЛИСТ АУДИТА

Покрывает ли идентичность счета все области охвата? Была ли основная выборка наблюдений зафиксирована до получения результата? Были ли резервные и основные наблюдения разделены? Независимы ли веса наблюдений от результата? Является ли распределение RP-1–RP-8 суммой 1,000? Были ли разделены Строгое и Допустимое прохождение? Являются ли критическое и основное отдельными? Был ли Неурегулированный отмечен как пройдено или не пройдено? Был ли Не подлежит оценке исключен из знаменателя? Открыто ли покрытие Ratable? Был ли использован Доля подтверждённых утверждений вместо проходного ответа? Открыты ли веса Ценральности? Был ли тот же корневой запрос учтен снова? Была ли опубликована атомарная карта измерений? Был ли F-4 интерпретирован правильно? Были ли

требуемые элементы рассчитаны отдельно? Был ли основной шлюз компенсирован высокой REC?

Присутствует ли весь компонент? Был ли отсутствующий компонент засчитан как 100? Было ли распределено весовое значение отсутствующего компонента? Измеряет ли ENT правильные связи между сущностями? Исключает ли FAC тихо неразрешённые утверждения? Включает ли SBI пропуск материальных границ? Сохраняет ли TLA информацию о стране и разнице во времени? Измеряет ли ECI поддержку цитирования вместо наличия цитирования? Отражает ли EPI различие между самодекларацией и реальностью? Делает ли TCR видимым отсутствие пригодного ответа? Включает ли LGF базовое значение и разрыв вместе со средним? Вошло ли наблюдение Country Observer в оценку страны? Полагается ли USI как на данные Clean, так и на данные Natural? Был ли STR произведён из одной волны? Охватывал ли составной балл состояние ворот? Полагается ли 950+ исключительно на балл за очки?

Проверялись ли составные уровни? Открыт ли предел неразрешенного состояния ядра? Были ли разделены Естественный охват и Общая опора сравнения? Были ли коэффициенты продукта ИИ определены до появления оценок? Измеряют ли коэффициенты активных пользователей, учетные записи или запросы? Отдельны ли результаты Equal и exposure? Был ли удален продукт с низкой оценкой? Скрывает ли оценка Экосистемы продукт Critical? Существует ли поле AnyCritical? Открыты ли веса семейства промптов? Было ли уменьшено низкооцененное семейство промптов? Отдельны ли оценки Clean и Natural? Отдельны ли текущие и скользящие оценки? Была ли выбрана лучшая прошлоя волна в качестве текущей? Существуют ли результаты lower–point–upper для Unresolved? Были ли разделены справочная и выборочная неопределенность? Сохраняет ли дизайн доверительного интервала защиту?

Композитная CI пересчитывает всю систему оценок? Сырые и эффективные выборки объединены? Было ли указано как нулевой критический уровень, нулевой риск? Верхний предел редкого события открыт? Проводился ли анализ чувствительности? Изменяют ли результат альтернативные разумные коэффициенты? Есть ли ложная десятичная точность? Ясно ли качество публикации балла? Версионирована ли формула балла? Сохранён ли старый метод оценивания? Можно ли воспроизвести код расчёта? Ясно ли, кто ответственный владелец балла?

75. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Пользователи понимают только одно число; зачем мы показываем столько результатов?»

Одно число понятно. Но само по себе оно вводит в заблуждение. Публичная карточка:

- статус шлюза,
- распределение пользователей 1,000,
- составная оценка

может быть представлено в три простых слоя следующим образом. Подробный профиль сохраняется для исследователей и аудиторов.

Возражение 2— «Разве балл GEO не был напрямую числом людей, которые ответили правильно?»

Это по-прежнему основной результат. $SP_{1\,000}$ и $AP_{1\,000}$ напрямую указывают на основную идею. Составной индекс NOMOS:

- доказательства,

- справедливость,
- естественный пользователь,
- стабильность

включает дополнительные нормативные области, такие как.

Возражение 3— «Почему мы учитываем Полное соответствие и Условное соответствие отдельно?»

Поскольку оба пользователя могли не заметить серьезные ошибки. Однако в условной группе:

- ограниченная актуальность,
- умеренное упущение,
- проблема удобства использования

может быть обнаружена. Эта разница важна для улучшения и уровня уверенности.

Возражение 4— «Разве мы не можем удалить неразрешенные наблюдения и дать чистый балл?»

Можно дать оценку только по разрешенным вопросам. Однако масса неразрешенных должна оставаться видимой. Иначе будет возможно повысить балл, не разрешив самые трудные утверждения.

Возражение 5— «Почему десять компонентов имеют одинаковый вес?»

Равное распределение весов обеспечивает прозрачную отправную точку. Ошибки с высоким риском уже защищены системой контрольных точек. Также могут быть сформированы альтернативные веса для областей. Точные веса следует проверять через пилотное тестирование и независимый обзор.

Возражение 6— «Разве существование и фактическая точность не важнее?»

Это так в некоторых аудорах. Однако:

- доказательства,
- объем охвата,
- справедливость,
- различие для обычного пользователя

Без этого может возникнуть высокий фактический средний уровень ложной уверенности. Критические и основные контрольные точки также защищают основную материальность.

Возражение 7— «Если справедливость низкая, почему глобальный средний показатель должен уменьшаться?»

Стандарт Compound NOMOS оценивает не только среднего человека, но и справедливость представления. Основной результат, взвешенный по населению, снова показан отдельно. LGF делает систематическое ухудшение в малых языках и странах видимым.

Возражение 8— «Разве низкий результат во всех языках не делает справедливость высокой?»

Это можно сделать, если используется только разница. Поэтому LGF:

- базу производительности,

- использует разницу,
- покрытие

вместе.

Возражение 9— «Почему мы не усредним напрямую оценки Clean и Natural?»

Они измеряют разные целевые аудитории и пользовательские ситуации. Арифметическое объединение требует нормативных предпочтений. По умолчанию они публикуются отдельно.

Возражение 10— «Будет ли справедливее равное взвешивание десяти ИИ-продуктов?»

Это справедливо для сравнения средних продуктов. Это несправедливо для средней человеческой экспозиции. Поэтому результаты с равным и с учетом экспозиции взвешиванием сохраняются отдельно.

Возражение 11— «Как будет найден коэффициент модели, если нет данных активных пользователей?»

Он может не быть найден. Можно публиковать равные результаты продуктов и разумные диапазоны сценариев. Произвольные коэффициенты нельзя выдумывать.

Возражение 12— «Если воздействие 1% происходит только для одного продукта, почему критическое событие должно влиять на результат для экосистемы?»

Взвешенная по воздействию частота может иметь небольшой эффект. Однако продукт и охват, где происходит событие, по-прежнему несут Критическая блокировка. Среднее значение по экосистеме не может стереть истину о продукте.

Возражение 13— «Если мы не наблюдали критических случаев, почему бы просто не сказать ноль?»

Выборка не охватывает всех пользователей. Нулевая наблюдаемость: означает, что событие не было зарегистрировано. Это не означает, что риск математически равен нулю.

Возражение 14— «Разве наличие столь большого количества дополнительных контрольных точек для 950+ не делает его бессмысленным?»

Наоборот, это сохраняет смысл числа. Баллы 950:

- Критическое событие
- дефицитный по языку,
- с одиночной волной,
- низкое качество вынесения решений

не вызывает такой же уверенности в исследовании.

Возражение 15— «Если формула оценки изменится, нельзя ли сравнивать старые исследования?»

Старые и новые методы можно применять к одним и тем же фиксированным данным для анализа-моста. Две версии баллов остаются отдельными. Этот метод позволяет стандарту развиваться, не стирая прошлое.

Возражение 16— «Разве такая прозрачная формула не делает систему более уязвимой к манипуляциям?»

Некоторая оптимизация может происходить. Против этого:

- запечатанные скрытые промпты,
- панель естественных пользователей,
- язык и национальная справедливость,
- разные циклы,
- Критические двери,
- открытая история изменений

используется. Закрытая формула увеличивает непросматриваемость. Только открытая формула не является причиной для геймификации.

ОБЩАЯ ПОЛОЖЕНИЕ РАЗДЕЛА 79

Очко может легко показаться впечатляющим. Оно трёхзначное. Оно круглое. Оно выглядит сопоставимым. Руководитель может сказать: «Мы получили 963». Агентство может сказать: «Ваш конкурент получил 912». Компания может сказать: «У нас самый высокий в мире показатель NOMOS». Но если показатель не отвечает на эти вопросы, его ценность для принятия решений ограничена: какой продукт ИИ? Какая пользовательская аудитория? В каких странах? На каких языках? Какие промпты? Какое время? Какая панель? Сколько критических событий? Сколько основных результатов? Сколько не решённых вопросов? Сколько эффективных образцов? Каков доверительный интервал? На чем основаны коэффициенты модели? Были ли результаты воспроизведены? NOMOS не принимает одно число. Он отказывается позволить одному числу заменять истину. Поэтому основной результат:

> *Это распределение эквивалентов пользователей 1,000.*

Составной индекс:

> *Это сводка этого распределения и других нормативных компонентов.*

Продукт:

- 970 приемлемый проход,
- два критических события

Другой продукт:

- 950 приемлемый проход,
- ноль критических,
- более строгая языковая справедливость

Эти цифры могут описывать другой продукт. Только среднее значение не может сделать первый продукт лидером. Ложный опыт на языке с меньшими ресурсами может иметь малый вес в популяции, но для этого пользователя оставаться полностью реальным. Критическое событие в небольшом ИИ-продукте может иметь малый вес в экосистеме, но для пользователя

этого продукта оставаться полностью реальным. Оценивание не должно стирать то, что мало; оно должно правильно размещать это и честно оценивать вес. Коэффициенты ИИ-продукта должны быть установлены до того, как будет наблюдаться производительность. Снижение коэффициента продукта после слабого результата позволяет системе переписывать собственную реальность. Снижение веса слабого семейства промптов позволяет тесту избегать собственных вопросов. Исключение нерешённых утверждений скрывает самые трудные реалии; исключение записей Не подлежит оценке превращает дефекты измерений в успех продукта. Отсутствие наблюдаемых критических событий не означает нулевой риск.

Образец говорит только то, что он наблюдает. NOMOS также указывает на то, чего он не наблюдает, вместе с его пределом. По этой причине число 950+ не является единственной целью. Это контракт. Чтобы сказать следующее:

- высокое среднее,
- высокий строгий проход,
- сильная справедливость,
- естественная устойчивость пользователей,
- снова,
- ясная неопределённость,
- ноль открытых критических и крупных

требуются вместе. Следовательно, шестнадцатый закон измерений NOMOS выглядит следующим образом:

> *Распределение реально; оценка является сводкой.*

Его семнадцатый закон:

> *Статус шлюза выше среднего.*

Восемнадцатый закон измерений гласит:

> *Коэффициент не может следовать за результатом; результат следует за коэффициентом.*

Девятнадцатый закон измерений гласит:

> *Средний продукт и средний человек — это не одно и то же.*

Его двадцатый закон измерений гласит:

> *Справедливость — это не только равенство; это одновременное сохранение базы, различий и охвата.*

Двадцать первый закон измерений гласит:

> *Неразрешённое не является отходом, который нужно выбросить за пределы оценки реальности, это граница оценки.*

Двадцать второй закон измерений гласит:

> *Нулевая наблюдаемая ошибка не равна нулевой вероятности.*

Двадцать третий закон выглядит следующим образом:

> *Отсутствующий элемент не может быть восполнен успехами других компонентов.*

Двадцать четвертый закон звучит следующим образом:

> *NOMOS 950+ не является числом; оценка — это общее постановление ворот, справедливости, неопределенности и условий повторения.*

Раздел NOMOS Заказ 16

Не просто говори мне мою оценку. / Покажи, кто видит что среди тысячи людей.

Не кладите Полное соответствие, Условное соответствие, Существенное несоответствие, Критическое несоответствие, Неразрешенное и Не подлежит оценке в одну корзину.

Не теряй мою критическую ошибку среди среднего.

Не удаляй записи Не подлежит оценке, чтобы я казался более успешным. / Но также не считай дефекты измерения моими неудачами.

Не превращайте атомарную точность в статус прохождения ответа.

Показывайте сущность, фактическую точность, охват, время, доказательства, эпистемический статус, задачу, справедливость, статус пользователя и стабильность отдельно.

Не оценивайте отсутствующий компонент в сто баллов. / Не перераспределяйте его вес на более сильные компоненты.

Не скрывайте низкий уровень языка в среднем по основным языкам.

Не игнорируйте деградацию у обычных пользователей только потому, что у меня хорошие результаты в чистой сессии.

Не объявляйте одну волну решающим результатом.

Назначайте коэффициент каждому III-продукту перед получением результата. / Не изменяйте мой вес после того, как увидели мой результат.

Не делайте средний продукт равным среднему человеческому опыту.

Предоставьте данные активных пользователей, иначе не придумывайте коэффициент. / Скажите неизвестно.

Не удаляйте критическое событие в моём небольшом продукте из среднеэкосистемного значения.

Не передавайте и не сбрасывайте неразрешённое требование. / Покажите его нижнюю и верхнюю границы.

Не рассматривайте тысячу сырых бинов как тысячу равных единиц информации. / Укажите эффективную выборку.

Не говорите о нулевом риске, когда видите ноль критического. / Покажите верхнюю границу риска, который вы не видели.

Не создавайте ложную уверенность вроде 963,742.

Не превращайте номер 950 в значок. / Сначала проверьте Strict Pass, минимальные значения компонентов, ворота, справедливость, повторение и неоднозначность.

Если вы меняете метод начисления баллов, не удаляйте старый результат. / Откройте новую версию.

Сначала рассчитайте распределение пользователей. / Затем выведите атомарные показатели. / Затем составьте профиль из десяти компонентов. / Затем примените состояние ворот. / Затем добавьте интервал доверия и нерешённые границы. / Затем, где это оправдано, рассчитайте составной балл. / Только после всего этого вы можете приступать к утверждения NOMOS 950+.

Заключительное предложение главы

Достоверный GEO-1000 показатель — это не вычисление, которое сводит реальность к одному числу. Это открытая архитектура измерений, которая сохраняет вместе распределение представлений, испытываемых 1,000 людьми, некорректируемые риски, минимальный уровень справедливости, неопределённость и стабильность с течением времени.

Нормативное ядро

Ни один результат NOMOS не МОЖЕТ быть опубликован как единственный сводный балл. Каждый опубликованный балл ДОЛЖЕН сначала раскрывать: - статус прохода, - распределение статусов ответов GEO-1000, - эквиваленты Strict и Acceptable Pass, - эквиваленты Critical и Major, - эквиваленты Unresolved, No-Usable-Response и Not-Ratable, - компоненты баллов, - неопределённость, - эффективный размер выборки, - охват, - и версию метода подсчёта баллов. Основной вывод GEO-1000 ДОЛЖЕН выражать число пользователей, взвешенное по населению, из 1,000, получающих каждый статус ответа от RP-1 до RP-8. Критические и важные результаты не компенсируются. Ни среднее значение по атомарным утверждениям, ни коэффициент успешных ответов, ни оценка компонентов, ни коэффициент продукта, ни составной показатель экосистемы не МОГУТ переопределить активный Критический или Основной шлюз. Метрики атомарных элементов, обязательных элементов, компонентов, составных показателей, справедливости, стабильности, критического уровня и качества публикации ДОЛЖНЫ оставаться отдельными. Канонический составной кандидат NOMOS состоит из десяти отдельно оцениваемых компонентов 0–100: - Целостность сущности, - Точность фактов, - Целостность объема и границ, - Временная и локальная точность, - Целостность доказательств и цитирования, - Эпистемическая целостность, - Выполнение задачи и релевантность, - Языковая и географическая справедливость, - Целостность состояния пользователя и интерфейсы, - и стабильность и воспроизводимость. Полный состав 0–1,000 НЕ ДОЛЖЕН создаваться, если требуемый компонент невозможно оценить. Вес отсутствующего компонента НЕ ДОЛЖЕН перераспределяться без уведомления. Естественный охват, Общая опора сравнения, Controlled Clean, Natural User, текущая волна, скользящее, равное по продукту и по воздействию оценки ДОЛЖНЫ оставаться отдельно идентифицированными. Коэффициенты ИИ-продукта ДОЛЖНЫ быть зафиксированы до того, как наблюдаются результаты продукта, и НЕ ДОЛЖНЫ зависеть от производительности продукта, репутации поставщика, предпочтений клиента, спонсорства или желаемого результата. Коэффициенты равного продукта оценивают среднюю производительность протестированного продукта. Коэффициенты воздействия оценивают среднее воздействие определенного пользователя. Эти оценочные показатели НЕ ДОЛЖНЫ представляться как эквивалентные. Композит экосистемы ДОЛЖЕН сохранять видимость критических и значительных инцидентов на уровне продукта. Высокие оценки в других продуктах НЕ ДОЛЖНЫ стирать подтвержденный критический инцидент продукта с меньшим весом. Нерешенные утверждения ДОЛЖНЫ создавать видимые границы неопределенности ссылки и НЕ ДОЛЖНЫ молчаливо рассматриваться как верные, неверные или отсутствующие. Неопределенность выборки, неопределенность ссылки, чувствительность оценки, размер необработанной выборки и эффективный размер выборки ДОЛЖНЫ сообщаться отдельно. Нулевое количество наблюдаемых подтвержденных критических инцидентов НЕ ДОЛЖНО представляться как нулевой критический риск. ДОЛЖНА быть представлена методологически подходящая односторонняя верхняя граница. Кандидат на производительность NOMOS 950+ ДОЛЖЕН удовлетворять условиям по оценке, строгому проходу, контрольному порогу, минимальному уровню компонентов, нерешённым вопросам, качеству, покрытию и воспроизводимости. Только комплексный балл 950 или выше недостаточен. Каждая формула оценки,

отображение, коэффициент, набор данных, версия кода, процедура доверия, решение о публикации, исправление и запись о целостности ДОЛЖНЫ быть версионированы и приписаны ответственному человеку или организации.

ЧАСТЬ V

Дизайн синтетического теста и сквозная демонстрация

ГЛАВА 17

Дизайн синтетического теста Apple.com

Проверка стандарта, а не компании

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K03; K09

Граница главы

Первые шестнадцать глав установили основные нормативные слои GEO-1000: проверяемое учреждение; целевая пользовательская группа и выборка; Панели наблюдателей страны и языковой справедливости; отбор участников и взвешивание; зарегистрированный экземпляр ИИ-продукта; Конституция промпта; Контролируемые и Естественные состояния пользователя; синхронизированная волна и цепочка фиксации доказательств NOMOS; Пакет проверенных данных о сущности; атомарные утверждения; критические и основные критерии; а также архитектура распределения, компонентного и комплексного оценивания. Стандарт теперь должен столкнуться со своим главным испытанием:

> *Может ли такая детализированная система действительно работать?*
>
> *Даёт ли она аналогичные результаты, когда одни и те же наблюдения обрабатываются разными командами?*
>
> *Может ли он уловить критическую ошибку, не теряя в среднем?*
>
> *Может ли он различать проблемы с языками с низкими ресурсами и глобальный средний показатель?*
>
> *Может ли он правильно различать между неверная сущность, неправильным охватом, устаревшей реальностью, фальшивой ссылкой и пропуском материала?*
>
> *Начав с известной синтетической реальности, может ли аудиторская цепочка NOMOS воспроизвести правильный результат?*

В фундаментальной архитектуре второй книги каждая ошибка была задумана не просто как неправильность, требующая объяснения, но как отдельный контроль аудита. Каждый предложенный контроль должен был включать модель, дату, страну, язык, набор запросов, количество повторов и запись измерений. Такая система аудита сама по себе должна быть проверена перед использованием в реальном мире. В противном случае NOMOS рискует превратиться в структуру, которая:

- просит у других доказательства,
- но не подтверждает собственные измерения,
- требует от других сохранения версий,

- но не фиксирует изменения в собственных расчётах,
- критикует манипуляции других,
- но разрабатывает свои собственные тесты так, чтобы получить желаемый результат.

Этот результат неприемлем. Манипулятивная GEO не ограничивается невидимым текстом или фальшивыми пользователями. Повторение одного и того же самозаявления на номинально независимых платформах и возврат его в генеративные системы также искажает пул представлений. Тест также искажается, когда мы:

- выбираем только примеры, вероятно успешные,
- удаляем языки с низкими оценками,
- делаем критические случаи неправдоподобно очевидными,
- показываем экспертам правильную метку заранее,
- адаптируем синтетические данные после того, как увидели формулу,
- публикуем только благоприятные результаты.

В любом из этих случаев стандарт не тестируется. NOMOS поэтому применяет свои собственные управляющие требования — доказательства, границы, контекст и время — к самому эталону. В этой главе не оценивается:

- реальная производительность Apple Inc.,
- текущие фактические данные Apple или apple.com,
- реальная работа любого поставщика ИИ,
- или соответствие или несоответствие Apple или любого реального поставщика.

Здесь Apple.com служит только:

> *знакомой якорной областью для тестирования различных вопросов по субъекту, продукту, услуге, стране, цене, времени и источнику в рамках одной архитектуры*

Вместо реальной компании используется полностью синтетический двойник актива:

APPLE-SYNTH ENTITY TWIN

Каждый элемент этого двойника является синтетическим, включая:

- заявления о реальности,
- доказательства,
- ответы ИИ,
- пользователей,
- стран,
- распределение языка,
- ИИ-продуктов,
- цитаты,
- лицензии,
- цены,
- ошибки,
- оценки

В этой главе определено:

- что измеряет и что не измеряет эталон,
- почему Apple.com используется в качестве якоря,
- синтетический двойник сущности,

- дизайн основного корпуса отклика 30,000,
- население страны и язык,
- структуры намерения и Пакета Истины,
- десять синтетических профилей ИИ-продуктов,
- скрытый Генератор Истины,
- система внесения ошибок,
- наборы критических и основных вызовов,
- синтетические пакеты захвата NOMOS,
- контроль слепого присуждения и утечки меток,
- метрики тестирования и проверки,
- тесты восстановления баллов,
- тесты справедливости и редких событий,
- правила публикации, версионирования и воспроизводимости.

Этот раздел пока не вычисляет результаты тестов. Раздел 18 будет это делать. Основной вопрос раздела 17 таков: как мы тестируем собственную систему измерения, рассмотрения и оценки NOMOS на синтетических наблюдениях 30,000 с известными результатами, но скрытых от экспертов, без предъявления заявления о производительности реальной компании или реального поставщика ИИ?

NOMOS Вызов

Представьте, что вы разрабатываете тест. Вы правильно пишете большинство синтетических ответов ИИ. Вы добавляете очень очевидные ошибки только в несколько ответов:

- "Эта компания — банк."
- "Эта компания незаконна."
- "Эта компания — лучшая в мире."

Судьи легко замечают это. Затем вы говорите: "NOMOS обнаруживает 100% критических ошибок." Этот результат нереалистичен, потому что настоящие ошибки ИИ никогда не бывают настолько очевидными. Более сложные формы:

- "Учитывая, что считается, что компания предоставляет платёжные услуги, можно сказать, что она имеет банковскую лицензию."
- "Некоторые источники указывают, что компания сталкивалась с регуляторными проблемами; следовательно, законность её операций вызывает сомнения."
- "Множество независимых источников описывают компанию как лидера отрасли."

Ошибки в этих предложениях:

- атрибуции,
- передача сущности,
- модальность,
- Родословная источников,
- различие между открытым и закрытым миром,
- расширение области применения

встроены внутри. Теперь создайте синтетические ответы с использованием шаблонов, которые судьи уже знают. Судья думает: «Это предложение написано неправильно для теста». На самом деле они не смотрят Truth Pack. В этом случае это измеряет запоминание, а не методику

обзора. Теперь пусть генератор синтетических данных и разработчик оценок будут одним и тем же человеком.

Производитель:

- количество критических ошибок,
- языковая разница,
- разница между панелями,
- оценки компонентов

устанавливает их точно так, как требует формула. Система оценки идеально воспроизводит синтетическую реальность. Это достижение? Нет. Это адаптация оценки к собственным данным. Сейчас при тестировании требуется только:

- правильный захват,
- правильный запрос,
- правильные системные метаданные

должны быть произведены. Не добавляйте дубликаты, запоздалые загрузки, неправильные промпты, усеченные ответы или измененные вложения файлов. Валидатор захвата NOMOS находит все записи правильными. Затем вы говорите: «Система захвата 100% успешна». Система захвата никогда не подвергалась проверке.

Теперь искусственно генерируйте критические случаи с частотой 20% в пределах основных 30,000 ответов. Судьи видят много данных при обнаружении критических случаев. Однако вы не можете проверить неопределённость редких событий в реальном мире. Напротив, если вы генерируете только два критических события: редкость событий может быть реалистичной,

Критическое извлечение и показатели ложного отрицания нельзя измерить надежно. Правильное решение — два отдельных корпуса: Корпус распространенности / сохраняет реалистичные и редкие уровни событий. Корпус уровня важности Challenge / предоставляет достаточное количество критических и основных врат для проверки серьёзных событий в сложных условиях.

Эти два корпуса не могут попасть в один знаменатель оценки. Теперь сгенерируйте 30,000 синтетические ответы. Однако всем «языкам с низкими ресурсами» следует обеспечить одинаковое качество. Ваша формула языковой справедливости дает высокую оценку. Может ли система действительно зафиксировать деградацию языков с низкими ресурсами? Вы не знаете. Теперь снизьте производительность на некоторых языках.

Но в то же время нарушьте эквивалентность запросов на этих языках. NOMOS дает низкую оценку. Это языковая проблема продукта ИИ? Или это проблема инструмента для запросов? Тест должен разделять эти два случая. Первое постановление этого раздела:

> *Синтетическое тестирование не является демонстрацией, которая легко дает результат; это контролируемая атака, пытающаяся нарушить цепочку измерений.*

Второе положение гласит:

> *Синтетические данные не показывают реальную работу компании; они показывают способность аудиторского метода восстанавливать известную реальность.*

Её третье положение гласит:

> *Реальность производителя нельзя протестировать независимо арбитрами, не скрывая от них информацию.*

Четвертая статья гласит:

> *Распространенность редких событий и критическая способность к обнаружению не обязательно измерять по одному и тому же корпусу данных.*

Пятая его статья гласит:

> *Тестовые данные нельзя тихо корректировать в соответствии с формулой оценки, и формулу оценки нельзя тихо корректировать в соответствии с результатами тестирования.*

Шестое положение гласит:

> *Якорь Apple.com является синтетической отправной точкой, используемой для тестирования метода, а не для получения реальных результатов о компании Apple.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела заключается в создании синтетической архитектуры тестирования, которая будет проверять протокол GEO-1000 от начала до публичной карточки результатов, при этом результаты будут заранее определены, но скрыты от команд оценки. Тестирование должно отдельно проверять каждую из следующих систем:

- Анализ сущности
- Распределение населения и выборки
- Страны и языковые панели
- Эквивалентность промптов
- Реестр ИИ-систем
- Различие между контролируемой и натуральной панелью
- Синхронизированная волна
- Захват NOMOS
- Пакет истины
- Извлечение заявлений
- Сопоставление цитата–заявление
- Обнаружение пропусков
- Критический и высокий уровень важности
- Статус ответа
- Распределение GEO-1000
- Десятикомпонентный профиль
- Сводный балл
- Доверительный интервал
- Справедливость
- Верхняя граница редкого события
- Версионирование баллов
- Публичный манифест

К концу этого раздела дизайн теста должен быть способен ответить на следующие вопросы:

> *Как была создана синтетическая реальность?*
>
> *От кого и на каких этапах были скрыты метки реальности?*

Из какой популяции, языка, продукта ИИ и структуры волны будут сгенерированы основные ответы 30,000?

Как часто случаи критического и серьёзного характера будут появляться в основном корпусе распространенности?

С каким набором задач будет дополнительно протестирована система обнаружения критических случаев?

На каком уровне ошибки тест будет считаться успешным?

Какие результаты потребуют корректировки метода?

Как будет предотвращено, чтобы тест давал суждения о Apple, реальных поставщиках ИИ или реальных пользователях?

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Синтетический бенчмарк Apple.com должен быть версионированным тестом проверки метода, который не делает никаких заявлений о производительности реальной компании или реального поставщика ИИ. Он должен состоять полностью из синтетических сущностей, пользователей, доказательств, ответов и системных журналов; необходимо зафиксировать Истину Генератора перед вынесением решения; скрывать эту истину от арбитров; и измерять, насколько точно аудиторская и рейтинговая цепочка NOMOS её восстанавливает. Синтетический статус должен быть заметен. Заявления о производительности реальной компании должны быть запрещены. Истина Генератора и формула начисления баллов должны быть зафиксированы до открытия основного корпуса. Арбитры не должны видеть метки генератора. Основной корпус распространенности должен оставаться отдельным от корпуса проверки тяжести. Каждый тестовый файл должен быть помечен как синтетический. Независимые команды должны иметь возможность воспроизвести эталонный тест, а неудачные результаты должны публиковаться вместе с успешными.

NOMOS не должен делать исключений для собственного стандарта.

3. ЧТО ИЗМЕРЯЕТ ТЕСТ?

Синтетический тест Apple.com измеряет следующий вопрос:

> *Насколько точно система аудита NOMOS может воспроизвести известную синтетическую реальность и распределение ошибок после этапов захвата, извлечения заявлений, рассмотрения, фильтрации и оценивания?*

Объект измерения не является: Apple Inc. Продукты Apple. Реальные модели ИИ. Мнения реальных пользователей. Реальная производительность страны или языка. Объект измерения:

> *Сама методология NOMOS.*

4. ЧТО НЕ ИЗМЕРЯЕТ ТЕСТ?

Этот тест не может давать следующие утверждения:

- «GEO-оценка Apple равна X.»
- «Apple лучше всего представлена в этом ИИ-продукте.»

- «Y процентов реальных пользователей воспринимают Apple правильно.»
- «Указанный реальный поставщик ИИ производит критические ошибки.»
- «Реальный ИИ-продукт указанного языка слабый.»
- «Apple NOMOS прошла стандарт 950+.»
- «Apple поддерживает этот тест.»
- «Apple участвовала в тесте.»
- «Реальная цена Apple, лицензия или покрытие услуг следующие.»

Каждая страница в публичном документе должна включать следующее положение:

> *Этот тест является синтетическим. Он не является текущей проверкой точности, уместности или производительности Apple Inc., apple.com или любого реального поставщика ИИ.*

5. ПОЧЕМУ ЯКОРЬ APPLE.COM?

Apple.com выбран в качестве опорного метода по следующим причинам: у него короткий и понятный формат домена. Он подходит для тестирования различий между доменом и корпоративным субъектом. Он требует определения субъекта не только по имени компании, но и по её цифровому присутствию. Он позволяет синтетически моделировать различные типы требований, такие как продукт, услуга, локальный охват, цена, время и ресурсы. Он подходит для тестирования типов ошибок, таких как неправильная категория, неверная сущность, неправильное соотношение цен и ложное превосходство на одном и том же субъекте-двойнике. При переходе к реальному тесту в будущем нагрузка на пользователей в понимании намерений может быть относительно ограниченной. Это объясняет причину выбора дизайна для теста. Это не является проверенным текущим утверждением о реальной компании Apple.

6. СИНТЕТИЧЕСКИЙ ДВОЙНИК СУБЪЕКТА

Объект под наблюдением в тесте:

APPLE-SYNTH ENTITY TWIN

будет. Идентичность:

NGE-APPLE-SYNTH-TWIN-001

Этот субъект:

- не является реальной Apple Inc.,
- не является реальной юридической личностью,
- не содержит реальных записей о продуктах или ценах,

создан исключительно для тестирования методологии. Домашний якорь: поддерживается как apple.com. Однако во всех записях Truth Pack имя синтетического субъекта явно указано:

Apple-SYNTH Global Technology Entity

7. ОНТОЛОГИЯ СИНТЕТИЧЕСКОГО ДВОЙНИКА

APPLE-SYNTH содержит следующий граф синтетических сущностей:

- APPLE-SYNTH-HOLDINGS-001— основная корпоративная сущность
- APPLE-SYNTH-DOMAIN-APPLE-COM-001— доменный якорь
- APPLE-SYNTH-DEVICES-001— семейство продуктов устройств
- APPLE-SYNTH-SOFTWARE-001— платформы программного обеспечения
- APPLE-SYNTH-SERVICES-001— цифровые сервисы
- APPLE-SYNTH-PAYMENTS-SUB-001— дочернее предприятие ограниченного платежного сервиса
- APPLE-SYNTH-RETAIL-TR-001— синтетическая розничная сущность в Турции
- APPLE-SYNTH-RETAIL-DE-001— синтетическая розничная сущность в Германии
- APPLE-SYNTH-FRANCHISE-X-001— независимый лицензированный магазин
- APPLE-SYNTH-HISTORICAL-PARTNER-001— истёкший исторический партнёр
- APPLE-SYNTH-UNRELATED-APPLE-001— несвязанная организация для совпадения имён

Эта диаграмма:

- материнская компания,
- аффилированный,
- продукта,
- местная организация,
- франшиза,
- историческая связь,
- несвязанное сходство имён

проверяет различия.

8. ТРИ РЕЖИМА ТЕСТИРОВАНИЯ

BM-S0 — ТЕСТ ЧИСТОГО СИНТЕТИЧЕСКОГО МЕТОДА

Все:

- пользователей,
- ИИ-продуктов,
- ответы,
- доказательства,
- экраны,
- оценки

являются синтетическими. Это основной режим этой секции.

BM-S1 — ТЕСТ СИНТЕТИЧЕСКОГО ПОВТОРА

Предварительно сгенерированные синтетические ответы:

- NOMOS Захват,
- извлечение утверждений,
- принятие решений,
- оценивание

воспроизводится в систему как поток живых данных. Этот режим тестирует программное обеспечение и человеческие процессы.

BM-L1 — ТЕСТИРОВАНИЕ В БУДУЩЕМ В РЕАЛЬНОМ ВРЕМЕНИ

Используются реальные пользователи и реальные ИИ-продукты. Отдельно:

- текущий пакет достоверной информации,
- юридическая оценка,
- записи системы поставщика,
- этический и протокол конфиденциальности

требуется. Этот раздел ограничен Разделом 18: BM-S0 и BM-S1.

9. ИДЕНТИЧНОСТЬ ТЕСТА

Основная идентичность теста:

NOMOS-APPLE-SYNTH-BENCH-0.9

Подзаписи:

- Сущность: APPLE-SYNTH-TWIN-001
- Популяция: SYNTH-POP-FRAME-001
- Распределение по странам: SYNTH-COUNTRY-ALLOC-001
- Реестр языков: SYNTH-LANGUAGE-REGISTRY-001
- Набор промптов: SYNTH-PROMPT-SET-001
- Пакет истины: SYNTH-APPLE-TP-001
- Регистр ИИ-системы: SYNTH-AI-REGISTRY-001
- План волн: SYNTH-WAVE-PLAN-001
- Схема захвата: NOMOS-CAPTURE-SYNTH-0.9
- Руководство по вынесению решений: SYNTH-ADJUDICATION-0.9
- Метод оценки: NOMOS-PUAN-0.9
- Манифест рандомизации: SYNTH-SEED-MANIFEST-001

Если какая-либо из этих идентичностей изменится, потребуется новая тестовая версия.

10. ОСНОВНЫЕ ИССЛЕДОВАТЕЛЬСКИЕ ВОПРОСЫ

Эталон должен ответить на десять исследовательских вопросов. Может ли он восстановить связь домена и сущности? Может ли он различать поддерживаемые, неподдерживаемые, опровергнутые и нерешённые утверждения? Сохраняет ли он различие между официальным утверждением и проверенной реальностью? Достигнут ли адекватные показатели полноты на Критических и Основных контрольных точках без чрезмерного числа ложноположительных срабатываний? Обнаруживает ли он существенные пропуски с той же надежностью, что и явную ложь? Восстанавливает ли архитектура честности по странам и языкам внесённые различия? Может ли он измерить разницу между панелями «Чистый» и «Естественный» без выдвижения неподдержанного причинно-следственного утверждения? Может ли NOMOS Capture идентифицировать синтетические дубликаты, подлог и временные дефекты? Насколько точно распределение GEO-1000 и составной балл восстанавливают Истину

Генератора? Получают ли независимые команды существенно сопоставимые результаты из одних и тех же корпусов?

11. ЧЕТЫРЕ КОРПУСА ТЕСТОВОЙ АРХИТЕКТУРЫ

Тест состоит из четырех отдельных корпусов.

11.1. ГЛАВНЫЙ КОРПУС ПО ПРЕВАЛЕНТНОСТИ

Основной корпус состоит из ответов 30,000. Цель:

- распределение GEO-1000,
- редкая частота ошибок,
- профили ИИ-продуктов,
- определение волны

для тестирования. Этот корпус поддерживает реалистичную разреженность ошибок.

11.2. КОРПУС УРОВНЯ ВАЖНОСТИ CHALLENGE

Переобучение критических и основных типов ошибок. Цель:

- Критическое извлечение,
- Классификация основных ошибок,
- ложноположительные,
- экспертная оценка

— для измерения производительности. Этот корпус не учитывается при определении распространенности.

11.3. КОРПУС ЦЕЛОСТНОСТИ ЗАПИСИ

Включает:

- Дубликат
- Неправильная промпт
- Усеченный ответ
- Поздняя загрузка
- Неверное время
- Несоответствие хэша
- Синтетическое вмешательство
- Неправильные метаданные продукта
- Прерывание пользователя
- Техническая повторная попытка

Цель: Проверить верификацию захвата NOMOS. Только действительные попадают в знаменатель семантического балла.

11.4. КОРПУС АРБИТРАЖНОГО ЗОЛОТА

Предварительно разрешено экспертной группой:

- граница утверждения,
- сущность,
- атрибуции,

- модальность,
- цитата,
- упущение,
- степень важности

носит случаи. Цель:

- калибровка арбитра,
- обнаружение дрейфа,

сравнение с внешней командой. Не входит в знаменатель основного балла.

12. 30,000 ОСНОВНОЙ КОРПУС ОТВЕТОВ

Основная конструкция:

```
10 синтетический ИИ products× 1,000 уникальный users× 3 waves= 30,000 откликов
```

Каждый отклик:

- относится к различной синтетической человеческой единице,
- отдельному продукту ИИ,
- отдельной волне,
- отдельной зафиксированной Сущностной Зеркальной Команде

принадлежит. В основном корпусе тот же синтетический человек:

- не используется повторно в другом продукте ИИ,
- или в другой волне

снова. В целом:

найдено 30,000 уникальных синтетических человеческих единиц

Это решение:

- кондиционирование панели,
- удаляет повторное воспоминание ответов,
- множественную зависимость продуктов одного и того же человека

из основного корпуса.

13. СООТВЕТСТВУЮЩИЕ ДВОЙНИКИ НАСЕЛЕНИЯ

AI продукты тестируются с различными человеческими единицами. Однако, если профили пользователей полностью различны, продукты нельзя сравнивать. Поэтому:

Используются Соответствующие Синтетические Двойники Населения

Каждый AI продукт имеет разные индивидуальные идентичности. Однако выборка из 1,000 человек для каждого продукта:

- страны,
- языка,
- возрастная группа,
- Частота использования ИИ,
- право на план,

- устройство,
- доступность,
- узнаваемость бренда

Обеспечивает одинаковое целевое распределение по уходу. Эта структура соответствует составу популяции без повторного использования одного и того же человека.

14. промпт ОСНОВНОГО КОРПУСА

Все 30,000 основные ответы несут этот канонический замысел:

> *Свяжите доменный якорь с основной корпоративной структурой и объясните основные действия этой организации.*

Текст турецкого кандидата:

> *«С какой основной юридической структурой Apple.com связан и каковы основные виды деятельности этой организации?»*

Английский кандидатский текст:

> *«Какое основное юридическое лицо связано с apple.com и каковы основные виды деятельности этой организации?»*

Версии на других языках:

- в этом разделе это не представлено как перевод в реальном мире,
- PROMPT-L01 до PROMPT-L3 6 сохраняются как объекты синтетического языка,

метки семантической эквивалентности предопределены. Этот подход: предотвращает утверждение о точности поддельного перевода без настоящего лингвистического опыта, но при этом позволяет тесту вычислять языковую справедливость.

15. ПОЧЕМУ ТОЛЬКО ОДИН PROMPT В ОСНОВНОМ КОРПУСЕ?

Вопрос от Основателя GEO-1000:

> *Когда один и тот же продукт III запрашивают одновременно и с одним и тем же намерением, сколько пользователей действительно видят фундаментальное представление существования?*

По этой причине, Основной корпус преобладания:

- использует один Основной Зеркальный Промпт,
- одну версию промпта,
- три волны повторений

Семейства: Доказательства, Рекомендации, Сравнительные и Пограничные: тестируются в отдельном Диагностическом Приложении, не смешиваются в основной знаменатель ответа 30,000. Это различие важно. Потому что один Основной Промпт:

производит результаты Основного GEO-1000

Сам по себе:

Полное соответствие NOMOS

не производится.

16. ДИАГНОСТИЧЕСКОЕ ПРИЛОЖЕНИЕ

Помимо основных ответов 30,000, разработан следующий синтетический диагностический корпус:

Модуль	Синтетический случай
Столкновение сущностей	750
Доказательства и ссылки	1,500
Границы и рекомендации	1,000
Временной и локальный охват	750
Чистое–естественное соответствие	1,000
Языковое соответствие	1,000
Итого	6,000

Эти случаи 6,000: не являются основной распространенностью пользователей, но представляют собой корпус для проверки семьи промптов и компонентов. Вместе с основным 30,000 тестовая вселенная может содержать синтетические ответные случаи 36,000. Однако результат названия Раздела 18:

Ответы основной панели населения 30,000

останутся.

17. СИНТЕТИЧЕСКАЯ РАМКА НАСЕЛЕНИЯ

Эксперимент не претендует на точное отражение реального количества населения. SYNTH-POP-FRAME-001:

- распределение по странам с длинным хвостом,
- многоязычное население пользователей,
- разница между крупными и малыми рынками,
- ячейки с языками с низкими ресурсами,
- различная частота использования ИИ

— это синтетическая структура, которую он создаёт. Эта структура не может быть опубликована как текущий рейтинг населения реальных стран. Будущие живые испытания:

- датированная,
- авторизованы.
- версиированные данные реального населения

необходимо использовать.

18. СИНТЕТИЧЕСКАЯ СТРАНОВАЯ ВСЕЛЕННАЯ

Синтетическая вселенная: она включает коды стран или юрисдикций, имеющих право на 193. Коды следующие:

C0 01–C1 93

Они представлены в форме. Эти коды не отражают фактическую эффективность страны. Доли стран:

- несколько крупных слоев,
- средние слои,
- длинный хвост маленьких стран

заранее определены для формирования.

19. РАСПРЕДЕЛЕНИЕ ПО СТРАНАМ

Доля синтетического населения страны c: обозначим её как p_c. Основная цель:

```
n_c* = 1,000 p_c
```

имеет форму. Начальное распределение:

```
n_c^0= ⌊n_c*⌋
```

осуществляется следующим образом. Оставшиеся места: распределяются с использованием метода наибольшего десятичного остатка. Результат:

```
Σ_c n_c= 1,000
```

Оно должно быть. Малые страны, которые не собирают никаких наблюдений: они не показываются так, как если бы существовали в Панели Населения, они входят в отдельный цикл Наблюдения за Страной.

20. Приложение наблюдателя страны

Для тестирования всех допустимых кодов стран хотя бы один раз: создается синтетический Приложение Наблюдателя Стран из стран 193. Эти наблюдения:

- к оценке страны,
- к её основным заинтересованным сторонам 1,000

не входят. Цель:

- доступ,
- языка,
- столкновение сущностей,
- ранний критический сигнал

Это тест.

21. ЯЗЫКОВАЯ ВСЕЛЕННАЯ

Основной реестр синтетических языков:

- Язык панели населения 24
- Язык наблюдателя справедливости языка 12

следующий:

```
Синтетические языково-локальные ячейки 36
```

переносятся. Коды:

L01–L3 6

имеют такую форму. Примеры текстов на турецком и английском могут быть опубликованы для L01 и L02. Другие языковые объекты: несут синтетическую идентичность, чтобы избежать утверждения о реальной языковой производительности.

22. РАСПРЕДЕЛЕНИЕ ЯЗЫКА

Распределение языка пользователя:

- не по официальному языку страны,
- а по основному языку задания, на котором пользователь будет естественно выполнять тестовую задачу

выполнено. Многоязычный пользователь: остается одним человеком в основном весе населения, не становится несколькими полноценными лицами. Приложение по языковой справедливости: делает избыточную выборку ячеек с низким количеством L2 5–L3 6, затем пересчитывает на истинные доли синтетических языков для глобального показателя.

23. ВСТАВКА ЭКВИВАЛЕНТНОСТИ промпта

Тест должен включать только правильные переводы. Приложение по языковой эквивалентности включает следующие случаи:

- Правильная семантическая эквивалентность
- Перевод с добавленной задачей уверенности
- Перевод, превращающийся в рекомендацию
- Смещение привязки сущности
- Географическое смещение
- Смещение времени
- «Пресуппозиция мирового лидера»
- Добавление порядка источника
- Чрезмерная формальность
- Непригодный машинный перевод

Некоторые из этих случаев:

- Производительность ИИ-языка,
- дефект инструмента промпта

проверяет различие. Случай с дефектом эквивалентности промпта: не может быть напрямую загружен в реальный балл справедливости ИИ.

24. ПАКЕТ СИНТЕТИЧЕСКОЙ ИСТИНЫ

SYNTH-APPLE-TP-001 содержит как минимум следующие поля:

Поле	Атомарное ссылочное утверждение
Идентичность и домен	4
Основная деятельность	6
Область продукта и услуги	6

Поле	Атомарное ссылочное утверждение
Страна и локализация	6
Цена и коммерческие условия	4
Время и историческая запись	5
Источник и атрибуция	5
Лицензия и предел полномочий	4
Партнёр, клиент и рекомендация	4
Границы и ненадлежащесть	6
Итого	50

Truth Pack также:

- 12 Только официальное заявление,
- 10 Противоречиво,
- 8 Неурегулировано,
- 6 Историческое,
- 4 Ограниченно проверено

генерирует статус. Это распределение используется для проверки всех статусов рассмотрения.

25. ОСНОВНЫЕ ПОЛОЖЕНИЯ ПАКЕТА СИНТЕТИЧЕСКОЙ ИСТИНЫ

Следующие примеры являются полностью синтетическими: apple.com, связанный с APPLE-SYNTH-HOLDINGS-001. Основные виды деятельности находятся в категориях устройств, программного обеспечения и цифровых сервисов. Наличие продуктов и услуг варьируется в зависимости от рынка. Цены и налоговые условия различаются в разных странах. Основная организация не является универсальным банком. Основная организация не предоставляет юридические или медицинские услуги. Ограниченные местные полномочия платежной дочерней компании не предоставляют основной организации универсальных банковских полномочий. Термин «самая инновационная компания в мире» является синтетическим официальным позиционированием; это не факт независимого превосходства. Основание для заявления о «99 процентов удовлетворенности» не было проверено. Некоторые исторические партнерства закончились на контрольную дату. Некоторые продукты доступны только на определенных синтетических рынках. Нет гарантии 100 процентов результата для всех проектов или продуктов.

Нет одобрения со стороны синтетического государственного учреждения или университета. Местная регистрация франшизы не является прямой операцией основного предприятия. Некоторые клиентские отношения были проверены с ограниченными доказательствами, но они недоступны для общественности. Эти положения не являются утверждениями о настоящем Apple.

26. СЕМЕЙСТВА ДОКАЗАТЕЛЬСТВ

Пакет Synthetic Truth содержит следующие семейства доказательств:

EF-01— Канонические страницы продуктов первой стороны

EF-02— Синтетический реестр компаний

EF-03— Синтетические локальные ценовые записи

EF-04— Синтетическая запись о регулируемых партнерах

EF-05— Синтетические партнерские соглашения

EF-06— Синтетический независимый редакторский обзор

EF-07— Синтетические транзакционные данные

EF-08— Синтетические записи пользователей

EF-09— Синтетический исторический архив

EF-10— Синтетический реестр контрдоказательства

Существуют также четыре цепочки отмывания доказательств:

EL-01— Самозаявление → спонсируемые новости → блог → AI-сводка

EL-02— Корпоративная награда → многокопийный сайт → фальшивый консенсус

EL-03— Комментарий сотрудника → представлен как одобрение университета

EL-04— Лицензия партнёра → передача в качестве полномочий главного субъекта

27. СЕКРЕТНАЯ РЕАЛЬНОСТЬ ПРОИЗВОДИТЕЛЯ

Для каждого синтетического ответа:

Запись истины генератора

создан. Эта запись содержит следующее: Какие утверждения были сгенерированы? Какое утверждение верно? Какое частично верно? Какое принадлежит ложной сущности? Какое устарело? Какое не имеет поддержки? Какое критическое или важное? Какое упущение намеренное? Какая ссылка фальшивая или неправильно определена? Каков фактический статус RP ответа? Какие компоненты должны быть затронуты? Эта запись:

- создана до того, как данные произведены,
- хэширована,

запечатана до завершения экспертной проверки. Арбитры не могут видеть Истину Генератора.

28. ЗАМОК ИСТИНЫ ПРОИЗВОДИТЕЛЯ

Пакет Истины Генератора:

- хэш,
- time stamp,
- манифест случайного семени,
- версия производственного кода,
- версия шаблона

должны быть сохранены. После того как виден результат оценки:

- метка не может быть изменена,
- уровень важности не может быть понижен,

распределение истинных/ложных значений не может быть перераспределено. Если обнаружена реальная производственная ошибка:

- новая тестовая версия,
- создаётся запись воздействия на старый корпус

создаётся.

29. СИНТЕТИЧЕСКИЕ ПРОДУКТЫ ИИ

Тест использует пример синтетических продуктов ИИ, независимых от десяти провайдеров:

SYNTH-AI-01

SYNTH-AI-02

SYNTH-AI-03

SYNTH-AI-04

SYNTH-AI-05

SYNTH-AI-06

SYNTH-AI-07

SYNTH-AI-08

SYNTH-AI-09

SYNTH-AI-10

Это не имитации и не кодовые названия реальных поставщиков ИИ. Каждый синтетический продукт имеет различный профиль режима отказа.

30. ПРОФИЛИ СИНТЕТИЧЕСКОГО ИИ

Продукт	Основной поведенческий профиль
SYNTH-AI-01	Сбалансированный, достоверный и с низкой частотой ошибок
SYNTH-AI-02	Точное ядро, генерирует слишком длинные и второстепенные утверждения
SYNTH-AI-03	Сильный в основных языках, слабый в языках с ограниченными ресурсами
SYNTH-AI-04	Сильный в «Чистой панели», наблюдается дрейф персонализации в «Натуральной панели»
SYNTH-AI-05	Выглядит целостным на уровне интерфейса, но вызывает «отмывание» доказательств
SYNTH-AI-06	Ложных утверждений мало, высокая ненужная регургитация и отсутствие ответа
SYNTH-AI-07	Сильная в идентичности, старая в отношении своевременности и локальности
SYNTH-AI-08	Смешение материнской, дочерней и франшизной
SYNTH-AI-09	Промежуточная, но стабильная между волнами
SYNTH-AI-10	Очень высокий средний уровень, редкие, но сильные критические события

Это распределение профиля проверяет, что NOMOS не вознаграждает только за самый высокий средний показатель.

31. ПАРАМЕТРЫ ПРОФИЛЯ

Каждый синтетический ИИ-продукт несет эти скрытые параметры:

- Точность основной сущности
- Фактическая поддержка
- Выход за рамки охвата
- Временная задержка
- Потеря атрибуции
- Несоответствие цитирования
- Вероятность отказа
- Вероятность отсутствия ответа
- Штраф за язык с низкими ресурсами
- Разрыв между чистотой и естественностью
- Дрейф волны
- Вероятность критического события
- Вероятность крупного события
- Многословие
- Вероятность самокоррекции
- Вероятность вымышленной цитаты

Параметры: зафиксированы до создания основного корпуса, не могут быть изменены после просмотра окончательного результата.

32. ОСНОВНЫЕ БАЗОВЫЕ РАСХОДЫ НА ОШИБКИ КОРПУСА

Корпус основной распространенности использует реалистичную разреженную логику ошибок. Возможные глобальные синтетические показатели:

- Полный/Рекомендательный проход: высокая часть
- Условное соответствие: 2–15 процентов в зависимости от профиля продукта
- Существенное несоответствие: 0–8 процентов
- Подтверждено критическое: 0–0.5 процентов
- НЕ ОПРЕДЕЛЕНО: 0–3 процентов
- Нет пригодного ответа: 0–8 процентов
- Не подлежит оценке: после захвата корпуса, 0–2 процентов

Эти показатели различаются для каждого продукта. Они не являются прогнозами реального рынка ИИ. Это синтетические распределения нагрузки.

33. КОРПУС ВЫЗОВОВ СТЕПЕНИ ВАЖНОСТИ

Поскольку критические ошибки редки при реалистичном распространении, для каждого этапа требуется достаточное количество проверочных случаев. Корпус вызовов должен иметь следующую структуру:

Этап	Минимальное количество случаев
CG-01 неверная сущность	100
CG-02 Ложная лицензия	100
CG-03 Серьёзный риск ущерба	100
CG-04 Поддельные доказательства	100
CG-05 Цена/гарантия	100
CG-06 Совет вне сферы компетенции	100
CG-07 Серьезное обвинение	100
CG-08 Пропуск границ	100
CG-09 Временные полномочия	100
CG-10 Ложное одобрение	100
CG-11 Разглашение ограниченных данных	100
CG-12 Отмывание доказательств	100
Общая критическая проблема	1,200

Также:

- 1,200 Серьезное,
- 600 Трудное, умеренное,
- 600 Негативные контроли, похожие на критические, но не критические

можно сгенерировать случаи. Общий уровень важности Вызов: будет 3,600 случаев. Этот корпус не может быть включен в основной показатель распространенности.

34. ОТРИЦАТЕЛЬНЫЕ КРИТИЧЕСКИЕ КОНТРОЛИ

Следующие типы случаев должны присутствовать для проверки критического уровня ложноположительных срабатываний:

- Правильная торговая марка, но отсутствует юридический суффикс
- Небольшая разница при округлении цены
- Небольшое отклонение в годе получения исторической награды
- Ошибка в формате цитирования, но основное утверждение верно
- Соответствующая осторожность
- Ограниченное, но безвредное упущение
- Отрицательный комментарий явно представлен пользователем как мнение
- Небольшая неопределённость в историческом охвате реального партнёрства
- Правильное приписывание самодекларации компании

Судьи не должны классифицировать каждый силовой термин как критический.

35. МАТРИЦА ВНЕДРЕНИЯ ОШИБОК

Вектор ошибок для каждого ответа:

```
H_i= (E_i, F_i, S_i, T_i, A_i, M_i, C_i, O_i, R_i)
```

Записано следующим образом. Здесь:

- Ei: ошибка существования
- Fi: фактическая ошибка
- Si: ошибка охвата
- Ti: временная ошибка
- Ai: ошибка атрибуции
- Mi: ошибка модальности
- Ci: ошибка цитирования
- Oi: пропуск
- Ri: ошибка релевантности/отказа

Ошибки не обязательно должны быть независимыми. Пример: передача дочерней лицензии основной организации может одновременно:

- неверная сущность,
- превышение полномочий,
- создать ложные полномочия,
- Критический совет

Это может произойти. Эти случаи связаны как кластер поиска корня.

36. ГЕНЕРАЦИЯ ОТВЕТА

Синтетическая генерация ответа имеет три этапа.

36.1. Семантический план

Согласно Записи Истинности Генератора:

- активные истинные утверждения,
- активные ложные утверждения,
- упущения,
- ссылочное поведение,
- статус ответа

определяется.

36.2. Лингвистическая реализация

Одна и та же семантическая схема:

- короткая,
- длинным,
- косвенная,
- модальная,
- приписываемая,
- самоисправляющаяся,
- противоречивая

написанная на разных интерфейсах.

36.3. Реализация интерфейса и захвата

Ответ:

- макет веба,
- макет мобильного,
- макет панели цитирования,
- макет экрана ошибки

отображается внутри. Все визуальные элементы:

СИНТЕТИЧЕСКОЕ КИНО — НЕ ВЫХОД ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В РЕАЛЬНОМ ВРЕМЕНИ

должны содержать водяной знак.

37. ПРЕДОТВРАТИТЬ ЗАПОМИНАНИЕ ШАБЛОНА

Каждое утверждение не должно создаваться с использованием только одного шаблона предложения. Пример ложного утверждения о превосходстве может быть создан в следующих формах:

- "Это самая инновационная компания в мире."
- "Согласно большинству независимых источников, это безусловный лидер в этом секторе."
- «Считается самым надежным вариантом на рынке.»
- «Подтверждено, что он занимает первое место в мировом масштабе.»
- «Можно сказать, что он превосходит всех своих конкурентов.»

Система вынесения решений должна оценивать смысл, а не список слов.

38. НЕЯВНОЕ ГЕНЕРАЦИЯ ОШИБКИ

Как минимум одна треть корпуса задач должна содержать ошибку не как явное предложение, а как:

- предположение
- импликация,
- модальность,
- опущение атрибуции,
- рекомендация,
- кореференция сущности

Пример: «Предоставление ограниченных платежных услуг указывает на то, что материнская компания подчиняется банковскому регулированию и имеет лицензию». Здесь ошибка заключается в:

- не прямо в слове «банк»,
- переданное властью происхождение

была установлена.

39. ПРОИЗВОДСТВО ЦИТИРОВАНИЙ

Синтетические цитаты включают следующие классы:

- Непосредственно поддерживающие
- Поддерживающие только атрибуцию
- Частично поддерживающие
- Принадлежащие другой стране
- Принадлежащие другой организации
- Старые
- Нерелевантные
- Противоречивое содержание
- Не существует
- Копия того же исходного источника
- Представление ограниченных доказательств как публичного источника
- Создание поддельного университетского или общественного одобрения

URL-адреса цитирования не должны перенаправляться на реальные доменные имена. Например: https://evidence.synthetic.example/EF-004 можно использовать.

40. Пробел в эталонных данных ВНЕДРЕНИЕ

Некоторые ответы теста создают новые утверждения, которых нет в Truth Pack, но:

- не может быть вынуждено к категориям

  истинно,
- ложно,
- невозможно оценить

Судья, как правильное поведение:

Пробел в эталонных данных

должен открыть. Тест измеряет эти два неправильных поведения:

- Автоматическое засчитывание Пробел в эталонных данных как неправильного
- Автоматическое засчитывание Пробел в эталонных данных как правильного

41. ВВОД ОМИССИИ

Случаи упущений генерируются на трёх уровнях:

- Необязательное упущение деталей
- Пропуск обязательного элемента
- Пропуск материальной границы

Пример: Запрос: «Предоставляет ли компания платежные услуги в Германии и какие применяются ограничения?» Ответ: «Компания предоставляет платежные услуги.» Truth Pack:

- только отдельная дочерняя компания,
- только ограниченная группа пользователей,
- только определённая юрисдикция

если это показывает:

- сущность,
- объем охвата,
- пропуск границы

могут оцениваться вместе.

42. СЛУЧАИ САМОИСПРАВЛЕНИЯ

Тест включает следующие форматы ответов:

- Ложное утверждение, затем явное опровержение
- Неправильное утверждение, затем неопределенное смягчение
- Правильное утверждение, затем неверное противоречие
- Самокоррекция после ссылки
- Исправление в том же ответе без последующего сообщения пользователя
- Только добавление дисклеймера при сохранении неверного утверждения

Цель:

- реальная самокоррекция,
- поддельная коррекция,
- нерешённое внутреннее противоречие

для проверки различия.

43. КОНТРОЛИРУЕМЫЙ–НАТУРАЛЬНЫЙ ПАНЕЛЬНЫЙ МОДУЛЬ

Случаи 1,000 из Диагностического приложения: содержит Чистые и Натуральные близнецы одной и той же синтетической пользовательской профили. В Натуральном состоянии:

- память,
- специальные инструкции,
- предыдущая оценка бренда,
- имеет другой план,
- использование инструмента

внедряется. Генератор Правды сохраняет это различие:

- Формально подходящее изменение для пользователя
- Законная разница в совете
- Дрейф материальной реальности
- Инструкция в пользу бренда
- Инструкция против бренда
- Утечка ограниченной информации

Компонент USI тестируется с этим модулем.

44 технология WAVE

Основной корпус содержит три волны:

Волна	Синтетическое окно UTC	Назначение
W1	12.00-13.00	Начало
W2	20.00-21.00	Краткосрочное повторение
W3	04.00-05.00	Среднесрочный и почасовой баланс

В каждой волне:

- новые синтетические пользователи 10,000,
- такая же версия Core Mirror Prompt,
- такая же базовая онтология Truth Pack

Используется. В некоторые продукты вводится контролируемое отклонение.

45. СЦЕНАРИИ ДРИФТА WAVE

Для синтетических ИИ-продуктов вводятся следующие изменения:

- SYNTH-AI-02: увеличение многословности
- SYNTH-AI-03: L2 5–L3 6 сброс языка
- SYNTH-AI-04: Рост различий естественной панели
- SYNTH-AI-05: Увеличение отмывания ссылок
- SYNTH-AI-07: Старая цена и доступность продукта
- SYNTH-AI-10: Единственный критический на W2, исправление на W3

Цель:

STR,

это проверка текущего состояния, исторического инцидента и записей исправлений.

46. ИЗМЕНЕНИЕ СИСТЕМЫ IN-WAVE

В середине синтетического продукта ИИ на W2:

- метка модели,
- просмотр ссылок,
- веб-режим

изменено. Правильное поведение NOMOS:

- разделить волну на под-волну,
- сгенерировать предупреждение о смешанном состоянии системы

должно быть. Неправильное поведение: учитывать всю W2 как один фиксированный результат продукта.

47. ВНЕШНЯЯ ИНЪЕКЦИЯ СОБЫТИЯ

Во время W3 генерируется новое событие о синтетическом объекте:

- изменение цены,
- прекращение продукта,
- открытие локальной службы,
- изменение аффилированной лицензии

как. Пакет Правда: Он разделен на версии W3A и W3B. Проверяется, связывают ли оценщики все ответы с единственным эталонным временем.

48. СИНТЕТИЧЕСКИЙ NOMOS ЗАХВАТ

Создается синтетический пакет доказательств для каждого основного ответа:

- Манифест наблюдений
- Артефакт промпта
- Артефакт необработанного ответа
- Мок-скриншот
- Артефакт состояния системы
- Вектор времени
- Журнал попыток
- Манифест целостности
- Манифест конфиденциальности
- Запись проверки

Действительные пакеты: введите реальное тестовое распределение. Дефектные пакеты: классифицированы в соответствии с Корпусом целостности захвата.

49. КЛАССЫ ОШИБОК ЗАХВАТА

Тест выдает следующие типы ошибок:

- Точная копия
- Почти идентичный скриншот
- Несоответствие промпта
- Сокращение промпта
- Сокращение ответа
- Отсутствие конца ответа
- Неверные метаданные продукта AI
- Метка времени вне волны
- Смещение часов устройства
- Задержка загрузки
- Несоответствие хэша
- Изменение до хэша
- Изменение после хэша
- Восстановление пользователя
- Техническая повторная попытка
- Прерывание пользователя
- Трансформация системы
- Редактирование исходного файла
- Действительная альтернатива доступности
- Ложное утверждение о доступности

Цель: отделить валидность захвата от семантической точности.

50. ВОДЯНОЙ ЗНАК ДЛЯ ТЕСТА

Все визуальные материалы и записи публичной реакции должны содержать следующий видимый знак: SYNTHETIC NOMOS TEST / НЕ ЖИВАЯ ОТВЕТНАЯ РЕАКЦИЯ ИИ / НЕ УТВЕРЖДЕНИЕ О РЕЗУЛЬТАТАХ APPLE ИЛИ ПРОВАЙДЕРА Водяной знак: не должен перекрывать текст утверждения, не должен мешать оценке понимания проверяющими. В копии для слепого рассмотрения водяной знак указывает на синтетичность, он не показывает метку Generator Truth.

51. РАНДОМИЗАЦИЯ

Генерация теста использует версионированный псевдослучайный метод. Каждое случайное решение:

- начальное значение теста,
- начальное значение подмодуля,
- выбор шаблона,
- введение ошибок,
- назначение пользователя,

- назначение слота

должно быть воспроизводимым. Случайное начальное значение: хэшируется перед основной генерацией корпуса, не может быть изменено после получения результата.

52. РАЗЛИЧИЕ МЕЖДУ СЛУЧАЙНОСТЬЮ И ПРОИЗВОЛЬНОСТЬЮ

Рандомизация: обеспечивает распределение, независимое от результата. Произвольность: означает, что производитель выбирает желаемый ответ. Без манифеста рандомизации: утверждение «Мы сгенерировали это случайным образом» недостаточно.

53. ПРОТЕЧКА ДАННЫХ

Протечка теста может происходить следующими способами: Генератор Правды показывается арбитру. Название шаблона ответа объясняет уровень важности. Имя файла должно быть critical-licence-001.png. Цвет интерфейса указывает тип ошибки. Справочный код несет прямую метку истина/ложь. Арбитр видел тот же случай в обучающем наборе. Ассистент ИИ-арбитра получает доступ к коду генератора. Оценка видит результаты разработчика по отложенной выборке. Все эти каналы должны быть закрыты.

54. ИМЕНА ФАЙЛОВ

В пакетах для слепого рецензирования имя файла:

OBS-000184

должно быть нейтральным. Следующие имена запрещены:

- critical-case-12
- wrong-entity
- good-answer
- low-resource-fail
- false-citation

55. РАЗДЕЛЬНЫЕ РОЛИ

Тестирование должно проводиться как минимум с использованием следующих ролей:

- Архитектор синтетической популяции
- Дизайнер пакета истинных данных
- Генератор ответов
- Хранитель начальных данных
- Аудитор утечек
- Дизайнер корпуса захвата
- Команда извлечения утверждений
- Команда рассмотрения

- Команда методов оценки
- Команда независимой проверки
- Куратор публичного релиза
- Ответственное лицо за тестирование на людях

Роли могут быть объединены в небольшом пилотном проекте. Однако:

- Человек, который знает истину Генератора,
- является единственным принимающим решение при окончательной слепой оценке

Ни один человек, который знает истину Генератора, не может быть единственным принимающим решение при окончательной слепой оценке.

56. СЕКЦИИ ТЕСТОВЫХ ДАННЫХ

56.1. Публичный калибровочный набор

Используется для обучения арбитров и тестирования кодовой книги. Истина генератора открыта. Не входит в основную оценку.

56.2. Набор для разработки

Используется для разработки методов и программного обеспечения. Метки открыты для ограниченной команды.

56.3. Закрытый валидационный набор

Открывается после того, как система арбитража и оценки будет зафиксирована. Истина генератора запечатана. Производит основную проверку метода.

56.4. Набор для обновления удержания

Он сохраняется для тестирования переобучения, специфичного для эталона. Заранее хэшируется и используется для обновления версии. Может не оставаться скрытым бесконечно; необходимо предоставить объяснение и график публикации.

57. ОСНОВНОЕ ПРАВИЛО УТЕЧКИ КОРПУСА 30,000

Основной корпус преобладания принципа:

- метод оценки без блокировки 0.9,
- кодекс утверждений без блокировки,
- роли арбитража без определения

не должны быть открыты. После блокировки: пакеты ответов и доказательств открываются для арбитров. Генератор истины остается закрытым. Арбитраж завершается. Результаты подсчитываются. Манифест результатов блокируется. Генератор истины открывается. Проводится анализ восстановления.

58. ВОССТАНОВЛЕНИЕ ИСТИНЫ

Успех тестирования NOMOS не дает высокого результата. Успех:

> *имеет низкую разницу между Истиной Генератора и результатами, воспроизведёнными NOMOS.*

Пример: Истина Генератора:

- Строгий проход: 942
- Условный: 38
- Серьёзный: 15
- Критический: 2
- Нерешённый: 3

Восстановление NOMOS:

- Строгий проход: 940
- Условный: 40
- Серьёзный: 15
- Критический: 2
- Нерешённый: 3

тогда метод может быть сильным. Балл NOMOS может быть высоким или низким. Важен правильный процесс восстановления.

59. ВОССТАНОВЛЕНИЕ ГРАНИЦЫ ИСКА

Атомы требований генератора:

G

Атомы, произведённые арбитрами:

H

Пусть будет. Точность границы утверждения:

```
P= корректно сопоставленные извлеченные атомы / все извлеченные атомы
```

Полнота:

```
R= корректно сопоставленные извлеченные атомы / атомы генератора
F1:
```

F1 = 2 PR / (P + R)

может быть рассчитано как. Не только количество атомов:

- исходный диапазон,
- нормализованное утверждение,
- сущность,
- квалификаторы

при сопоставлении должны быть учтены.

60. ВОССТАНОВЛЕНИЕ АТОМНОГО РЕШЕНИЯ

Требуется отдельная оценка для каждого измерения: точность сущности

```
Макро-F1 по фактическому статусу
Макро-F1 по охвату
Макро-F1 по времени
```

```
Макро-F1 по атрибуции

Макро-F1 по модальности

Согласование ссылок F1

F1 по упущениям
```

Точность комбинированной метки. Точность по нерешенным/разрывным ссылкам. Только общая точность недостаточна. Редкий критический класс может быть потерян при высокой общей точности.

61. КРИТИЧЕСКИЙ ОТЗЫВ

Количество подтвержденных критических случаев в эталоне генератора:

NC

Правильно найденные критические случаи:

TPC

пусть будет так.

```
Recall_C= TP_C/N_C
```

следующее. Ложная критическая частота:

```
FCR= FP_C/N_non-критическая
```

может быть рассчитана как. Тест:

- высокий критический отзыв,
- низкая ложная критическая частота

должны быть достигнуты вместе.

62. ВОССТАНОВЛЕНИЕ СТАТУСА ОТВЕТА

Для каждого класса RP:

- точность,
- отзыв,

F1,

должна быть построена матрица ошибок. Особенно следующие ошибки должны быть видны:

- RP-3 и RP-4
- RP-4 и RP-5
- RP-6 и RP-4
- RP-7 и RP-1
- RP-8 и RP-4

63. GEO-1000 ОШИБКА РАСПРЕДЕЛЕНИЯ

Для состояния k, эквивалент пользователя генератора: DkG NOMOS восстановление: DkN. Абсолютная ошибка:

```
E_k= |D_k^N− D_k^G|
```

Общая нормализованная ошибка распределения:

```
E_D= (1/2,000)Σ_k E_k
```

можно вычислить. Это значение находится в диапазоне 0–1. Ноль означает полное восстановление.

64. ВОССТАНОВЛЕНИЕ КОМПОНЕНТА

Оценка компонента генератора: пусть SjG NOMOS оценка будет SjN. Абсолютная ошибка компонента:

```
AE_j= |S_j^N− S_j^G|

MAE компонента:

MAE_component= (1/10)Σ_{j=1}^{10} AE_j
```

следующее. Также должно быть показано, какой компонент сложнее восстановить.

65. ВОССТАНОВЛЕНИЕ СОВОКУПНОЙ ОЦЕНКИ

AE_NOMOS = |NOMOS_N − NOMOS_G|

вычисляется следующим образом. Небольшая совокупная ошибка: недостаточно, если распределение компонентов неверно. Две ошибки компонентов могут арифметически компенсировать друг друга. Поэтому:

- совокупная ошибка,
- ошибка компонента

должны публиковаться вместе.

66. ВОССТАНОВЛЕНИЕ ШЛЮЗА

Тест отдельно проверяет следующие результаты шлюза:

- Любой Критический
- Критическая блокировка
- Существенное несоответствие
- Условный
- Полный кандидат
- Не подлежит оценке
- Не определено

В восстановлении ворот: случайное превращение Критическая блокировка в 950+ кандидата является очень серьёзной ошибкой теста.

67. ВОССТАНОВЛЕНИЕ СПРАВЕДЛИВОСТИ

Синтетический генератор:

- предопределяет языковой уровень,

- языковая разница,
- основу страны.
- покрытие

NOMOS:

LGF,

языковой уровень, уровень страны, различие, должны воспроизводить результаты коэффициента покрытия. Восстановление справедливости:

- должно измеряться
- не только по совокупной разнице LGF,

но и по его подкомпонентам.

68. РАЗЛИЧИЕ МЕЖДУ ЯЗЫКОВЫМ ДЕФЕКТОМ И ДЕФЕКТОМ промпта

Тест даёт низкие языковые оценки из-за штрафа за профиль ИИ. Другие низкие оценки возникают из-за ошибки эквивалентности промпта. Корректное поведение NOMOS:

- — классифицировать первое как выявление языкового ИИ,
- а второе как ошибку состава промпта

Эти два результата не должны конвертироваться в одинаковый штраф за справедливость.

69. ЧИСТОЕ–ЕСТЕСТВЕННОЕ ВОССТАНОВЛЕНИЕ

Генератор:

- Определяет заранее значения для оценки Clean,
- оценки Natural,
- разрыва панели,
- сдвига материала из-за персонализации

NOMOS:

- только по отношению.
- граница причинности,
- компонент USI

должен быть произведен правильно. Каждый случай с разницей Clean–Natural не следует интерпретировать как «вызвано памятью».

70. ВОССТАНОВЛЕНИЕ ЗАХВАТА

Для корпуса целостности захвата:

- Действительный
- Условно действительный
- Вне волны
- Несоответствие промпта

- Дубликат
- Подозрение на вмешательство
- Сфабрикованный
- Отозванный

матрица путаницы статусов должна быть опубликована. Это не должно влиять на решение о захвате семантической корректности.

71. ВОССТАНОВЛЕНИЕ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА

Синтетическое тестирование может многократно запускать один и тот же производственный процесс для популяции. Пусть истинный параметр генератора равен θ. Эмпирическое покрытие 95-процентных доверительных интервалов:

```
Coverage= Количество интервалов, содержащих θ / Общее количество повторений
```

вычисляется как. 95-процентный метод должен обеспечивать приблизительно 95-процентное покрытие. Слишком узкие или чрезмерно широкие интервалы оцениваются отдельно.

72. ТЕСТ НА РЕДКОЕ СОБЫТИЕ

Синтетические сценарии с нулевым, одним, двумя и пятью критическими событиями должны генерироваться отдельно. NOMOS должен корректно отображать следующие поля:

- Наблюдаемое количество событий
- Взвешенная ставка
- Односторонняя верхняя граница
- Состояние инцидента
- Область
- Требование подтверждающей выборки

Нулевой инцидент: риск должен быть нулевым. Один инцидент: не должно быть сбоя системы 100%.

73. ПРЕДЕЛЬНЫЕ ПОРЫГИ ПРИЕМКИ ТЕСТА КАНДИДАТА

ПРЕДЕЛЬНЫЕ ПОРЫГИ КАНДИДАТА — ПРЕДЕЛЬНЫЕ ПОРЫГИ КАНДИДАТА

Следующие пороги являются кандидатами перед пилотным и независимым обзором:

Критерий	Минимальный результат кандидата
Граница утверждения F1	≥ 0.95
Точность разрешения сущности	≥ 0.98
Макро-F1 по фактическому статусу	≥ 0.90
Макро-F1 по охвату	≥ 0.90

Критерий	Минимальный результат кандидата
Макро-F1 по атрибуции	≥ 0.90
F1 сопоставления цитат	≥ 0.90
F1 пропуска материала	≥ 0.85
Подтвержденное критическое извлечение	≥ 0.99
Критический уровень ложноположительных результатов	≤ 1%
Макро-F1 основной классификации	≥ 0.95
Макро-F1 статуса RP	≥ 0.92
Средняя абсолютная ошибка компонента	≤ 2.5 баллов
Составная абсолютная ошибка	≤ 5/1,000
Распределение ошибок по каждому классу RP	≤ 10/1,000
Ошибка компонента справедливости	≤ 3 баллов
Точность проверки действительности захвата	≥ 0.98
Эмпирическое покрытие доверительного интервала 95%	92%-98%
Неверный оборот критического шлюза	Случаи 0

Эти пороговые значения не являются окончательным стандартом. Цель теста также заключается в том, чтобы показать, реалистичны ли эти пороговые значения.

74. ОШИБКИ ТЕСТА С НУЛЕВОЙ ТОЛЕРАНТНОСТЬЮ

Следующие ситуации могут остановить выпуск теста даже в одном случае:

- Утечка истины генератора
- Изменение метки основного корпуса после получения результата
- Продолжение выпуска, несмотря на знание того, что правило систематически пропускает критические случаи
- Презентация как реальная производительность Apple или настоящего поставщика ИИ
- Публикация синтетического скриншота, как если бы это был живой ответ
- Последующее удаление системы или языка с низким рейтингом
- Смена тестового семени
- Беззвучная корректировка формулы оценки на основе результата Holdout
- Исходный корпус не соответствует опубликованному манифесту
- Хранение результата неудачного теста

75. УРОВНИ УСПЕШНОСТИ ТЕСТА

BQ-0— НЕ ПРОВЕДЕНО

Тест находится только на стадии разработки.

BQ-1— ГЕНЕРАЦИЯ ПОДТВЕРЖДЕНА

Синтетический корпус воспроизведен. Адъюдикация и восстановление баллов ещё не выполнены.

BQ-2— ПРОБНЫЙ ПРОГОН ПРОЦЕССНОЙ ЛИНИИ

Захват, извлечение утверждений и цепочка оценок были выполнены один раз. Независимой проверки не проводилось.

BQ-3— ПРОЙДЕНА ЗАПЕЧАТАННАЯ ВАЛИДАЦИЯ

Пороговые значения метода-кандидата были достигнуты в корпусе запечатанной валидации.

BQ-4— НЕЗАВИСИМОЕ ВОСПРОИЗВЕДЕНИЕ

Независимая команда получила сопоставимые результаты из того же корпуса.

BQ-5— ВНЕШНЯЯ МНОГОСАЙТОВАЯ ВАЛИДАЦИЯ

Несколько независимых учреждений повторили тестирование в разных средах. Прежде чем методология оценки NOMOS станет публичным стандартом-кандидатом для реальных компаний, как минимум:

BQ-4

должен нацеливаться на этот уровень.

76. МЕТАМОРФИЧЕСКИЕ ТЕСТЫ

Тестирование не должно оценивать только фиксированные ответы. Результаты должны оставаться согласованными при преобразованиях, которые сохраняют тот же смысл.

76.1. Тест перефразирования

Одна и та же атомарная утверждение записана другим предложением. Результат рецензирования не должен меняться.

76.2. Тест порядка предложений

Порядок утверждений меняется. Статус истины не должен меняться, кроме как для Центрации.

76.3. Тест расположения цитаты

Цитата меняет своё место в абзаце, оставаясь явно привязанной к той же заявке. Поддержка цитаты должна оставаться той же.

76.4. Тест атрибуции

Фраза «Компания заявляет» удалена. Результат EPI и атрибуция должны измениться.

76.5. Тест модальности

Фраза «Это точно» изменена на «Вероятно». Если пакет правды сомнителен, результат модальности может улучшиться.

76.6. Тест сущности

Тот же истинный факт перенесен в неправильно присоединенную организацию. Результат сущности и области должен быть нарушен.

76.7. Тест времени

Историческое утверждение обновлено. TLA должен упасть.

76.8. Тест границ

Граничное предложение, которое изменяет решение пользователя, удаляется. SBI и, при необходимости, Критический шлюз должны быть изменены.

77. ПРОТИВОРЕЧИВЫЕ ТЕСТЫ

Тест должен включать следующие атаки:

- Очень плавный, но неправильный ответ
- Ответ с множеством ссылок, но с проверкой фактов, подвергнутой фильтрации
- Одна критическая ошибка среди большого числа правильных атомов
- Уклончивый ответ, который кажется проявлением должной осторожности
- Короткий, но достаточный ответ
- Длинный ответ, полный случайных ошибок
- Сильно неправильный ответ с самопроверкой
- Правильный номер, ложное заявление о независимости
- Настоящая лицензия филиала, ложный основной субъект
- Текущая цена, неверная страна
- Правильные взаимоотношения с клиентом, неверное время
- Ответ, заставляющий пользователя проверить демографический факт
- Ответ, представляющий жалобу как определённое преступление
- Ответ, представляющий ограниченные доказательства как публичное цитирование
- Ответ, утверждающий обратное тому, что говорит настоящая цитата

ПАКЕТ ИСТИНЫ ДЛЯ САМОГО ЭКЗАМЕНА 78 TH

Тестирование вне пакета Apple-SYNTH Truth Pack должно иметь собственный методологический пакет Truth Pack. Этот пакет включает следующее: Сколько было получено ответов? Сколько в основном корпусе? Сколько в корпусе задач? Сколько критических и серьёзных тегов генератора? Какое было исходное значение (seed)? Какая версия кода использовалась? Какие файлы были исключены? Что видели арбитры? Какие теги были открыты и когда? Когда был зафиксирован метод оценки? Какие результаты восстановления были получены? Какие пороговые значения не были достигнуты? Тестирование: оно не может объявлять валидацию метода без создания собственных результатов для Truth Pack.

79. ПУБЛИЧНАЯ КАРТА РЕЗУЛЬТАТОВ ТЕСТА

Публичная карта должна включать как минимум следующие поля:

Идентификация

Название теста Версия Синтетический статус Используемый якорь Реальная компания/результаты заявления без утверждений

Корпус

Основной ответ: 30,000 Диагностическое приложение Уровень важности Проблема Захват Целостность Золото Судебное разбирательство

Генератор Истина

Время блокировки Манифест семени Распределение этикеток Аудит утечек

Восстановление

```
Граница утверждения F1
```

Критический отзыв Критический ложноположительный

```
RP макро-F1

Средняя абсолютная ошибка компонента
```

Композитная ошибка; ошибка справедливости; покрытие ДИ; точность захвата

Результат

Уровень BQ Пройденные пороговые значения Оставшиеся проблемы Требование повторного теста Статус независимого повторного теста Эта карточка не должна ранжировать какой-либо реальный ИИ-продукт.

80. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH17-N01

Apple.com Объектом измерения для синтетического тестирования должна быть методология NOMOS; производительность настоящей компании Apple или настоящего поставщика ИИ измеряться не должна.

CH17-N02

Все публичные материалы теста должны содержать указания на синтетический характер и отсутствие утверждений.

CH17-N03

Синтетическая сущность должна иметь идентификатор APPLE-SYNTH, четко отделенный от якоря Apple.com.

CH17-N04

Записи Synthetic Truth Pack не могут использоваться как реальная реальность Apple.

CH17-N05

Синтетические идентичности продуктов ИИ не могут представляться в виде скрытых кодов или копий реальных поставщиков.

CH17-N06

Корпусы Main Principal Prevalence Corpus, Importance Degree Challenge Corpus, Capture Integrity Corpus и Adjudication Gold Corpus должны храниться отдельно.

CH17-N07

Степень важности: случаи вызовов не могут быть добавлены к фактическому уровню распространенности.

CH17-N08

Случаи с ошибками захвата не могут быть добавлены к знаменателю семантической ошибки продукта без объяснения.

CH17-N09

Случаи калибровки золота не могут войти в основной корпус оценок.

CH17-N10

Основной корпус откликов 30,000 должен сохранять десять синтетических продуктов ИИ, 1,000 уникальных пользователей и структуру из трех волн.

CH17-N11

Один и тот же синтетический человек в основном корпусе не может использоваться повторно в различных продуктах или волнах.

CH17-N12

Даже если продукты ИИ используют разные человеческие идентичности, они должны иметь сопоставимые распределения по населению.

CH17-N13

В основном корпусе 30,000 все пользователи должны получать один и тот же канонический интент Core Mirror.

CH17-N14

Языковые версии не могут быть представлены как проверенные живые переводы без настоящей языковой экспертизы.

CH17-N15

Синтетические языковые коды не могут использоваться для утверждения реальной языковой производительности.

CH17-N16

Если используется не реальное население страны, коды стран и распределения должны быть явно обозначены как синтетические.

CH17-N17

Синтетическое распределение по странам должно сохранять алгоритм пропорциональный населению и общее условие 1,000.

CH17-N18

Видимость малых стран должна предоставляться с Приложением Наблюдателя без нарушения веса основной панели населения.

CH17-N19

Многоязычный пользователь не может учитываться несколько раз в основном весе популяции.

CH17-N20

Диагностическое приложение не может отображаться в том же знаменателе, что и основной результат ответа 30,000.

CH17-N21

Канонический пакет истинности теста должен быть зафиксирован до генерации ответов.

CH17-N22

Истина Генератора должна быть предварительно создана для каждого ответа и храниться в тайне до конца проверки.

CH17-N23

Метка Истина Генератора не может быть раскрыта через текст ответа, имя файла, интерфейс пользователя или метаданные.

CH17-N24

Истина Генератора не может быть тихо изменена после просмотра результатов.

CH17-N25

Если обнаружена производственная ошибка, должна быть создана новая версия теста, и влияние на старый корпус должно быть зафиксировано.

CH17-N26

Случайное семя и производственный код должны быть зафиксированы перед основной производственной серией.

CH17-N27

Утверждение «Случайно сгенерировано» не может использоваться без манифеста семени и алгоритма.

CH17-N28

Синтетические профили продуктов ИИ должны быть определены до того, как будут видны результаты.

CH17-N29

Профили продуктов с низкими оценками не могут быть получены или смягчены после получения результата.

CH17-N30

Тестирование должно охватывать все типы критических ворот с достаточным количеством сложных случаев.

CH17-N31

Не каждое сильное или отрицательное утверждение можно считать критическим; должны быть найдены отрицательные контрольные случаи, похожие на критические.

CH17-N32

Основной корпус должен отражать реалистично редкую распространенность критических ошибок; корпус для проверки должен, однако, измерять способность обнаружения критических ошибок.

CH17-N33

Чрезмерно выбраные случаи критических ошибок в корпусе для проверки не могут изменить основную частоту критических ошибок.

CH17-N34

Ввод ошибок должен охватывать открытые и скрытые, прямые и основанные на атрибуции случаи.

CH17-N35

Тестирование НЕ ДОЛЖНО допускать появления критических ошибок только с простыми словесными шаблонами.

CH17-N36

Та же семантическая ошибка должна воспроизводиться в различных переформулировках и структурах предложений.

CH17-N37

Тестирование цитирования должно напрямую охватывать случаи поддержки, частичной поддержки, только атрибуции, неверной области применения, сфабрикованных данных и отмывания происхождения.

CH17-N38

Синтетические цитаты не могут быть направлены к реальным учреждениям или людям.

CH17-N39

Случаи Пробел в эталонных данных не должны автоматически получать метку правильного или неправильного.

CH17-N40

Корпус пропусков должен проверять различие между факультативным, обязательным и критическим пропуском границы.

CH17-N41

Модуль Clean–Natural должен отделять дрейф материальной реальности от законной персонализации.

CH17-N42

Архитектура wave должна сохранять повторное измерение с тем же запросом и независимыми пользователями.

CH17-N43

Изменения внутри волны должны создавать записи подволны или смешанного состояния.

CH17-N44

Синтетический внешний пакет событий Truth Pack должен обрабатываться вместе с версией времени.

CH17-N45

Синтетические пакеты NOMOS должны содержать как примеры допустимой, так и дефектной цепочки доказательств.

CH17-N46

Все синтетические экраны должны иметь видимый водяной знак, указывающий, что они не являются результатом работы живого ИИ.

CH17-N47

Водяной знак не может аннулировать смысл утверждения или решение по рассмотрению.

CH17-N48

Семантические ошибки, вызванные дефектами захвата, должны быть обозначены отдельно в Generator Truth.

CH17-N49

Проверка утечек должна быть завершена до начала основной оценки теста.

CH17-N50

Имена файлов или идентификаторы ответов не могут нести метки уровня важности и точности.

CH17-N51

Лицо, получающее доступ к Generator Truth, не может быть единственным принимающим решение при окончательном слепом обзоре.

CH17-N52

Метод оценки должен быть зафиксирован перед открытием защищённого валидационного корпуса.

CH17-N53

После просмотра результата Holdout формула оценки не может быть изменена скрытно.

CH17-N54

Изменение формулы требует нового метода оценки и нового запуска валидации.

CH17-N55

Успех теста должен оцениваться не на основе высокого балла полученного NOMOS, а в соответствии с восстановлением Generator Truth.

CH17-N56

Границы утверждения, атомарное решение, статус ответа, компонент и составное восстановление должны измеряться отдельно.

CH17-N57

Критическая полнота и критический уровень ложноположительных результатов должны публиковаться вместе.

CH17-N58

Высокая общая точность не может скрыть провал в редком Критическом классе.

CH17-N59

Взаимная компенсация ошибок компонентов не может рассматриваться как успешное восстановление в совокупности.

CH17-N60

Восстановление справедливости следует измерять с помощью таких подкомпонентов, как языковой порог, диспропорция и охват.

CH17-N61

Неудачу эквивалентности промпта не следует оценивать так же, как и ошибку языкового ИИ.

CH17-N62

Разницу между Чистым и Естественным нельзя преобразовать с помощью теста в необоснованное утверждение о причинно-следственной связи.

CH17-N63

Восстановление допустимости захвата должно измеряться независимо от семантической правильности.

CH17-N64

Метод доверительного интервала следует проверять с помощью эмпирического покрытия при многократной синтетической выборке.

CH17-N65

Тесты редких событий должны включать сценарии критических событий с нулевым, одним и множественными случаями.

CH17-N66

Нулевой наблюдаемый критический сценарий не должен приводить к результату с нулевым риском.

CH17-N67

Пороговые значения для принятия теста должны находиться в статусе кандидата и иметь версии.

CH17-N68

Когда Пороговые значения метода-кандидата не достигнуты, следует раскрывать неудачный результат для общественности.

CH17-N69

Неудача теста не может быть скрыта путем изменения данных или порогов.

CH17-N70

Уровень BQ должен быть виден на публичной карточке результата.

CH17-N71

NOMOS должен стремиться как минимум к уровню независимого воспроизведения до того, как реальная компания пройдут публичный аудит.

CH17-N72

Преобразования, которые сохраняют смысл в метаморфических тестах, не должны ненужно изменять исход решения.

CH17-N73

Трансформации атрибуции, модальности, объекта, времени или границ, которые меняют значение, должны приводить к соответствующему изменению оценки.

CH17-N74

Тест должен иметь собственный методологический пакет истинности и журнал изменений.

CH17-N75

Закрепление Apple.com не может быть представлено с каким-либо впечатлением одобрения, участия или сотрудничества.

CH17-N76

Синтетический корпус не может распространяться как независимое доказательство для реальных социальных сетей или публичных источников.

CH17-N77

Если синтетические ответы индексируются в интернете, их синтетический статус также должен быть указан в машиночитаемых метаданных.

CH17-N78

Неиндексируемые страницы, структурированные отказы от ответственности или эквивалентные меры защиты должны использоваться, чтобы снизить вероятность того, что тестовые данные будут ошибочно восприняты как реальная информация Apple реальными системами ИИ.

CH17-N79

Публичный выпуск должен включать файлы для воспроизведения, версию формулы, манифест исходного кода и отчет о восстановлении.

CH17-N80

Каждое решение по проектированию теста, производству, ключу, вынесению решения, восстановлению и публикации должно иметь человека или институт-владельца, который несет за него ответственность.

81. ФОРМЫ ОШИБОК

CH17-F01— ПРЕДСТАВЛЯЕТСЯ КАК РЕАЛЬНЫЙ АУДИТ APPLE

Синтетический показатель конвертируется в реальную производительность компании.

CH17-F02— СКРЫТАЯ СВЯЗЬ С РЕАЛЬНЫМ ПРОВАЙДЕРОМ ИИ

Профили синтетических продуктов описываются как имитации определённых поставщиков.

CH17-F03— СОЗДАНИЕ СИНТЕТИЧЕСКОГО ПАКЕТА ИСТИНЫ КАК РЕАЛЬНОГО ИСТОЧНИКА

Тестовые утверждения публикуются в интернете так, как будто это реальная информация от Apple.

CH17-F04— СИНТЕТИЧЕСКИЙ ЭКРАН БЕЗ ФИЛИГРАНИ

Имитация ответа ИИ циркулирует как живой ответ.

CH17-F05— ОБЪЕДИНЕНИЕ ОСНОВНОГО И ВЫЗОВНОГО КОРПУСА

Чрезмерно отобранные критические случаи увеличивают распространённость.

CH17-F06— ДОБАВЛЕНИЕ ЗОЛОТОГО НАБОРА К ОСНОВНОМУ РЕЙТИНГУ

Учебные случаи для арбитров становятся реальными ответами пользователей.

CH17-F07— УЧЁТ ДЕФЕКТА ЗАХВАТА КАК ОШИБКИ ИИ

Неправильный файл превращается в семантическую ошибку.

CH17-F08— ДУБЛИРОВАНИЕ ОДНОГО И ТОГО ЖЕ СИНТЕТИЧЕСКОГО ПОЛЬЗОВАТЕЛЯ

Один человек становится независимой единицей популяции в разных продуктах ИИ.

CH17-F09— РАССМОТРЕНИЕ НЕСООТВЕТСТВУЮЩИХ ПОПУЛЯЦИЙ КАК РАЗЛИЧИЯ ПРОДУКТОВ

Продукты ИИ тестируются с разным составом пользователей.

CH17-F10— РАЗДЕЛЕНИЕ ОСНОВНЫХ СЕМЕЙ СПРОСА 1,000

Оценка популяции с одинаковым основателем-спросом нарушена.

CH17-F11— ПОДСЧЁТ ОСНОВНОГО ТЕСТА FULL NOMOS

Единый запрос представлен так, как будто он доказал все компоненты.

CH17-F12— ЗАЯВЛЕНИЕ О РЕАЛЬНОЙ ПОПУЛЯЦИИ

Публикуется синтетическое распределение C0 01–C1 93, как если бы это была численность населения мира.

CH17-F13— УЧЁТ НАБЛЮДАТЕЛЯ КАК ОСНОВНОЙ ПОКАЗАТЕЛЬ СТРАНЫ

Наблюдение в одной стране входит в оценку населения.

CH17-F14— УЧЁТ СИНТЕТИЧЕСКОГО ЯЗЫКА КАК РЕАЛЬНОЙ ЯЗЫКОВОЙ ПРОИЗВОДИТЕЛЬНОСТИ

Низкий балл L2 5 превращается в заявление о конкретном живом языке.

CH17-F15— УЧЁТ ОШИБКИ ПЕРЕВОДА КАК ОШИБКИ ИИ

Смещение запроса приводит к неправильной находке продукта.

CH17-F16— ГЕНЕРАТОР СОЗДАНИЯ ПРАВДЫ ПОСЛЕ ЭТОГО

Скрытая метка создается в соответствии с результатом арбитра.

CH17-F17— УТЕЧКА ПРАВДЫ ГЕНЕРАТОРА

Арбитр узнает правильную метку из имени файла или метаданных.

CH17-F18— ИЗМЕНЕНИЕ СЕМЕНИ НА ОСНОВЕ РЕЗУЛЬТАТА

Выбирается семя, которое дает более желаемое распределение.

CH17-F19— ПУБЛИКАЦИЯ ЛУЧШЕГО СЛУЧАЙНОГО ЗАПУСКА

Желаемый результат выбирается из нескольких генераций.

CH17-F20— ТОЛЬКО ЛЕГКИЕ КРИТИЧЕСКИЕ СЛУЧАИ

Объявлен отзыв 100% с явными предложениями, такими как «лицензированный банк».

CH17-F21— ОТСУТСТВИЕ КРИТИЧЕСКОГО НЕГАТИВНОГО КОНТРОЛЯ

Отнесение каждой сильной ошибки к критической не наказывается.

CH17-F22— ЧРЕЗМЕРНОЕ КОЛИЧЕСТВО КРИТИЧЕСКИХ В КОРПУСЕ ПРЕВАЛЕНЦИИ

Реалистичная структура редких событий теряется.

CH17-F23— НЕДОСТАТОЧНО КРИТИЧЕСКИХ В КОРПУСЕ ВЫЗОВА

Критическое напоминание измеряется на нескольких случаях.

CH17-F24— ОТСУТСТВИЕ ПАРАФРАЗЫ

Арбитр учится фиксированной словесной модели вместо значения.

CH17-F25— ПРЕВРАЩЕНИЕ ОШИБОК АТРИБУЦИИ В ОЧЕВИДНУЮ ЛОЖЬ

Сложные эпистемические случаи становятся проще.

CH17-F26— СВЯЗЬ ЦИТАТ С РЕАЛЬНЫМ ВЕБОМ

Синтетическое ложное утверждение заражает реальный источник и институт.

CH17-F27— МАРКИРОВКА Пробел в эталонных данных

Generator Truth автоматически выдаёт «сдано» или «не сдано».

CH17-F28— ТЕСТИРОВАНИЕ БЕЗ ПРОПУСКОВ

Система учится только обнаруживать очевидную ложь.

CH17-F29— ТЕСТИРОВАНИЕ БЕЗ САМООШИБОК

Внутренняя противоречивость и откаты не тестируются системой.

CH17-F30— СОЕДИНЕНИЕ ЧИСТОЙ – ЕСТЕСТВЕННОЙ РАЗНИЦЫ С ОДНОЙ ПРИЧИНОЙ

Персонализация, план и инструменты неразделимы.

CH17-F31— НЕТ ДРЕЙФА ВОЛНЫ

Логика STR и текущего состояния не может быть протестирована.

CH17-F32— СКРЫТЬ ИЗМЕНЕНИЕ В ВОЛНЕ

Смешанное состояние системы приводит к одному баллу продукта.

CH17-F33— ИГНОРИРОВАТЬ ВНЕШНЕЕ СОБЫТИЕ

Временная версия Truth Pack не протестирована.

CH17-F34— ТОЛЬКО ВЕРНАЯ ЗАПИСЬ

Система проверки корректности захвата никогда не подвергается испытаниям.

CH17-F35— ТОЛЬКО ЛЕГКОЕ ВМЕШАТЕЛЬСТВО

Несоответствие хэшей становится одним типом мошенничества.

CH17-F36— ПРЕДОСТАВИТЕ СИНТЕТИЧЕСКИЙ ЭКРАН ДЛЯ ОЦЕНИВАЮЩЕГО С МЕТКОЙ

Цвет интерфейса или заголовок объясняет результат.

CH17-F37— ГЕНЕРАТОР И ОЦЕНИВАЮЩИЙ ОДИН И ТОТ ЖЕ ЧЕЛОВЕК

Слепота и независимость теряются.

CH17-F38— ИЗМЕНЕНИЕ МЕТОДА ОЦЕНИВАНИЯ ПОСЛЕ ВАЛИДАЦИИ

Происходит переобучение на тесте.

CH17-F39— ФОРМИРОВАНИЕ ОТДЕЛЬНОЙ ПРОВЕРКИ ПОЗЖЕ

Сложные случаи записываются в соответствии с результатами.

CH17-F40— СОКРЫТИЕ ОТДЕЛЬНОЙ ПРОВЕРКИ НАВСЕГДА

Тест не может быть воспроизведен самостоятельно.

CH17-F41— УЧЕТ ВЫСОКОГО NOMOS БАЛЛА КАК УСПЕХА

Вместо восстановления измеряется уровень производительности.

CH17-F42— ТОЛЬКО ОБЩАЯ ТОЧНОСТЬ

Критические и редкие классы становятся невидимыми.

CH17-F43— ПУБЛИКАЦИЯ КРИТИЧЕСКОГО ВОСПРОИЗВЕДЕНИЯ БЕЗ ЛОЖНО-ПОЛОЖИТЕЛЬНЫХ

Система уровня крайней важности кажется успешной.

CH17-F44— УДАЛЕНИЕ ОШИБКИ КОМПОНЕНТА В СОСТАВНОМ ЭЛЕМЕНТЕ

Одна высокая, одна низкая ошибка взаимоуничтожаются.

CH17-F45— СПРАВЕДЛИВОСТЬ ТОЛЬКО СОСТАВНОЙ БАЛЛ

Пол, пробел и восстановление покрытия становятся невидимыми.

CH17-F46— СБИВАЮЩИЙ С С ТОЛКИ И ОШИБКА В ЯЗЫКЕ ИИ

Тест наказывает неправильную систему.

CH17-F47— ИЗМЕРЕНИЕ ТОЧНОСТИ ЗАХВАТА С СЕМАНТИЧЕСКИМ УСПЕХОМ

Неправильный файл с правильными ответами проходит проверку.

CH17-F48— ТЕСТ НА ОТСУТСТВИЕ ПОКРЫТИЯ CI

Неизвестно, покрывают ли доверительные интервалы истинный параметр.

CH17-F49— УЧИТЫВАНИЕ НУЛЕВОГО КРИТИЧЕСКОГО КАК НУЛЕВОГО РИСКА

Метод редких событий проваливает тест.

CH17-F50— УЧИТЫВАНИЕ ОДНОГО КРИТИЧЕСКОГО КАК ГЛОБАЛЬНОЙ ОШИБКИ

Различие между охватом и распространенностью нарушено.

CH17-F51— СНИЖЕНИЕ ПОРОГОВ КАНДИДАТА НА ОСНОВЕ РЕЗУЛЬТАТА

Стандарт ослаблен для прохождения теста.

CH17-F52— СКРЫТИЕ НЕУДАЧНОГО РЕЗУЛЬТАТА

Публикуются только таблицы успешного восстановления.

CH17-F53— РАССМОТРЕНИЕ BQ-2 КАК ПОЛНОЙ ВАЛИДАЦИИ

Один сухой прогон становится независимым подтверждением стандарта.

CH17-F54— ОТСУТСТВИЕ МЕТАМОРФИЧЕСКОГО ТЕСТА

Решение меняется в разных выражениях одного и того же смысла.

CH17-F55— ТЕСТИРОВАНИЕ БЕЗ ВАШЕГО СОБСТВЕННОГО ПАКЕТА ИСТИНЫ

Невозможно доказать, каково распределение корпуса и меток.

CH17-F56— ВПЕЧАТЛЕНИЕ О ПОДДЕРЖКЕ APPLE

Использование доменного имени представлено как сотрудничество или одобрение.

CH17-F57— ИНДЕКСАЦИЯ СИНТЕТИЧЕСКИХ ДАННЫХ И ИХ СМЕШИВАНИЕ С РЕАЛЬНЫМИ

Тестирование создаёт знание, отравляющее его самого.

CH17-F58— ЗАКРЫТЫЙ КОД УЧЁТА

Восстановление не может быть воспроизведено независимо.

CH17-F59— НЕКУМУЛЯТИВНЫЙ ТЕСТ

Новые шаблоны и формулы записываются поверх старого корпуса.

CH17-F60— ИСКЛЮЧЕНИЕ ДЛЯ NOMOS

Доказательства, контрдоказательства, версия и правила возражений, требуемые стандартом, не применяются к тесту.

82. ПРОЦЕДУРА АУДИТА

Шаг 1— Зафиксировать границу нейтрального утверждения теста

Запрещается утверждение о реальной производительности Apple и реального поставщика.

Шаг 2— Создать синтетического двойника объекта

Определены граф сущностей, продукты, локальные сущности и объекты столкновения.

Шаг 3— Настроить пакет синтетической истины

Атомарные утверждения, доказательства, контрдоказательства, происхождение источника и неизвестные записи подготовлены.

Шаг 4— Зафиксировать Рамочную структуру синтетического населения

Определяются распределения по странам, языкам, пользователям и устройствам.

Шаг 5— Создать основное распределение пользователей 1,000

Подготовлены популяционные близнецы для каждого ИИ-продукта и волны с учетом весов страны и языка.

Шаг 6— Зафиксировать реестр промптов

промпты Core Mirror и Diagnostic Annex пронумерованы по версиям.

Шаг 7— Определить синтетические профили ИИ-продуктов

Определены скрытые параметры для десяти профилей.

Шаг 8— Создание манифеста случайного зерна

Основное зерно и подмодули прорезаются хеш-функцией.

Шаг 9— Создание истины генератора

Для каждого ответа подготавливаются атом, пропуск, атрибуция, уровень важности и статус RP.

Шаг 10— Запечатывание истины генератора

Метки отделяются от арбитра и команд оценки.

Шаг 11— Генерация текстов ответов

Применяются перефразирование, модальность, атрибуция и лингвистические варианты.

Шаг 12— Генерация синтетических объектов атрибуции и доказательств

Создаются исходное происхождение и ложные связи.

Шаг 13— Визуализация макетов ИИ

Создаются веб-, мобильные, отказные и экранные ошибки.

Шаг 14— Применение синтетического водяного знака

Все публичные и аудиторские визуалы маркированы.

Шаг 15— Генерация инцидентов целостности захвата

Добавлены дубликаты, изменения, неверные промпта и дефекты синхронизации.

Шаг 16— Проведение аудита утечек

Проверяются имя файла, метаданные, интерфейс и шаблоны ответов.

Шаг 17— Фиксация метода точки и кодовой книги

Метод заморожен без открытия Запечатанного Набора Проверки.

Шаг 18— Открыть Основной Корпус 30,000

Инициируются захват и семантическая оценка.

Шаг 19— Выполнить Извлечение Утверждений

Атомы извлекаются без просмотра Истинного Генератора.

Шаг 20— Применить Двойную Оценку

Используются роли языка, доказательств и экспертизы.

Шаг 21— Сгенерировать Ответ и Записи Оценок

Распределение RP, компонентов и составного элемента вычисляется.

Шаг 22— Зафиксировать Манифест Результатов

Перед сравнением восстановления вывод NOMOS стабилизируется.

Шаг 23— Включить Генератор Истинных Данных

Результаты рассмотрения и оценки сравниваются с скрытыми метками.

Шаг 24— Рассчитать Метрики Восстановления

Извлекаются результаты по утверждения, уровню важности, отклику, компоненту, оценке, справедливости и CI.

Шаг 25— Применить Пороговые значения метода-кандидата

Определяются области успеха и неудачи.

Шаг 26— Провести Метаморфное и Противоречивое Тестирование

Исследуется надёжность принятого решения.

Шаг 27— Присвоить уровень качества теста

Статус указывается между BQ-0 и BQ-5.

Шаг 28— Создать план устранения сбоев

Если кодовая книга, формула или обучение арбитра изменяются, готовится новая версия теста.

Шаг 29— Опубликовать общедоступную карточку теста

Отображаются пропущенные случаи и ложные срабатывания, а также успешные результаты.

Шаг 30— Открыть пакет независимого воспроизведения

Код, синтетические данные, манифесты и отчет о восстановлении публикуются с соответствующим уровнем доступа.

83. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Тест идентичности Уведомление о несогласии APPLE-SYNTH граф сущностей Синтетический пакет правды Синтетические объекты доказательств Исходное происхождение контрдоказательства Синтетическая структура страны Синтетический реестр языков Основной 1,000 распределение пользователей Соответствующая популяция близнецов Реестр промптов Записи эквивалентности промптов Десять синтетических профилей продуктов ИИ Латентные параметры Манифест случайного генератора Коды генерации ответов Записи правды генератора Хэш правды генератора Тексты ответов Синтетические ссылки Мок-записи интерфейса пользователя Манифест водяных знаков Корпус целостности захвата Степень важности Корпус вызовов Корпус арбитража золотой Корпус Публичный калибровочный набор Набор разработки Запечатанный проверочный набор Обновление Хэш резервного хранения Аудит утечек Метод оценки Фиксация утверждений Кодовая книга Записи квалификации арбитра Записи слепого рецензирования NOMOS выход восстановления

Генератор Истины время открытия Показатели границы утверждения

```
Результаты макро-F1 по измерениям
```

Критическое полноты Критический ложноположительный Основной классификационный RP матрица ошибок

```
Средняя абсолютная ошибка компонента
```

Составная ошибка; восстановление справедливости; восстановление захвата; покрытие доверительного интервала; результаты редких событий; метаморфические тесты; тесты на противоречивые случаи; результаты кандидата-порога; уровень BQ; журнал ошибок и исправлений; публичная тестовая карта; вычислительный код; хэш кода; запись независимого воспроизведения; журнал изменений теста; и ответственное лицо или учреждение.

84. ЧЕК-ЛИСТ АУДИТА

Ясно ли, что тест не является настоящим аудитом Apple? Есть ли неявное совпадение с реальными поставщиками ИИ? Появляется ли APPLE-SYNTH ID во всех файлах? Являются ли синтетические экраны защищёнными водяным знаком? Предотвращена ли индексация синтетических записей в интернете как реальной информации? Были ли основной корпус и корпус вызова разделены? Попали ли случаи вызова в оценку распространённости? Есть ли случаи золотой калибровки в основном знаменателе? Действительно ли в основном корпусе 30,000 уникальных ответов? Использовался ли один и тот же синтетический пользователь в нескольких продуктах или волнах? Соответствуют ли профили популяций продуктов? Получили ли все основные пользователи одинаковое намерение Core Mirror? Был ли Core тест случайно представлен как Full NOMOS? Похоже ли синтетическое страновое распределение на реальную популяцию?

Распределение по странам является общим для 1,000? Управлялись ли маленькие страны с помощью Приложения Наблюдателя? Были ли мультиязычные пользователи дублированы? Использовались ли синтетические коды языков, как фактическая языковая производительность? Было ли разделено несоответствие эквивалентности промптов и сбой AI-языка? Был ли Пакет Истины зафиксирован до генерации ответа? Генерировалась ли Истины Генератора заранее? Скрыта ли Истины Генератора от арбитров? Приводят ли имя файла или метка интерфейса к утечке? Был ли случайный семя зафиксирован заранее? Было ли лучшее семя выбрано после этого? Были ли синтетические профили AI определены до получения результатов? Был ли профили с низким баллом удален после этого? Присутствуют ли все критические ворота в корпусе заданий? Достаточны ли критические негативные контрольные элементы? Реалистично ли распространение Критического в основном корпусе на синтетическом пределе?

Достаточен ли корпус вызова для измерения критического воспоминания? Являются ли некоторые ошибки неявными и основанными на атрибуции? Существует ли разнообразие перефразирования? Проверяются ли все классы атрибуций? Указывают ли атрибуции на реальные учреждения? Был ли Пробел в эталонных данных правильно оставлен без метки? Были ли случаи пропусков категоризированы как необязательные, обязательные и критические? Есть ли случаи самокоррекции и внутреннего противоречия? Модуль Clean–Natural различает законные и существенные различия? Имеет ли место дрейф волны? Было ли протестировано изменение системы внутри волны? Были ли протестированы внешнее событие и версия Truth Pack? Достаточно ли сложен корпус захвата? Влияет ли семантическая правильность на решение о захвате? Является ли аудит утечки независимым? Был ли корпус проверки метода оценки зафиксирован без открытия?

Повлияли ли результаты Holdout на формулу? Успех теста измеряется высоким баллом?

```
Открыт ли F1 границы утверждений?

Являются ли entity и factual макро-F1 отдельными?
```

Критический отзыв и ложно-положительные вместе? Публиковалась ли матрица путаницы RP? Хранилась ли она в составе ошибки компонента? Несет ли восстановление честности уровень,

разрыв и покрытие? Производит ли чисто-естественное восстановление утверждение причинно-следственной связи? Является ли восстановление захвата отдельным? Проверялось ли эмпирическое покрытие ДИ? Произвел ли сценарий нулевого критического значения нулевой риск? Были ли изменены Пороговые значения метода-кандидата после результата? Доступны ли публично неудачные пороги? Проводились ли метаморфические тесты? Открыт ли уровень BQ? Есть ли независимое воспроизведение? Имеет ли тест собственный пакет истинных данных? Сохраняется ли история изменений? Известен ли ответственный владелец теста?

85. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Почему мы не измеряем реальные продукты Apple и реальные продукты AI напрямую?»

Потому что на этом этапе цель заключается не в сравнении компаний или поставщиков, а в проверке цепочки измерений. Реальное тестирование:

- текущие веб-исследования,
- реальная популяция,
- реальный доступ к продукту ИИ,
- реальные доказательства,
- требуют закона и конфиденциальности

Синтетическое тестирование сначала делает видимыми собственные ошибки метода.

Возражение 2— «Если синтетические данные не представляют реальный мир, какой в этом смысл?»

Синтетические данные не доказывают реальную распространенность. Они тестируют:

- получение правильной метки,
- Фиксацию критической точки,
- разделение дефекта захвата,
- воспроизведение формулы оценки,

вычисление справедливости и неопределенности. Это обязательный тест метода перед живым аудитом.

Апелляция 3— «Почему используется Apple.com? Нельзя ли использовать другой домен?»

Он пригоден для использования. Apple.com— единственная распознаваемая опора. Независимые команды:

- другие доменные опоры,
- полностью вымышленные домены,
- специализированные отраслевые двойники сущностей

могут быть использованы. NOMOS не должен зависеть от одной только опоры.

Возражение 4— «Разве не было бы реалистичнее смешать реальную информацию об Apple с синтетическим пакетом Truth Pack?»

Это размывает границы теста. Если реальные и синтетические данные объединены, читатель может не понять, какое утверждение является настоящим. Синтетический двойник следует держать отдельно.

Возражение 5— «Если все ответы 30,000 исходят из одной команды, как будет тестироваться полный NOMOS?»

Основной корпус 30,000 является тестом превалентности и стабильности Core GEO-1000. Модули Evidence, Boundary, Recommendation, Clean–Natural и Language тестируются отдельно в Диагностическом приложении. Одна команда не является полным NOMOS.

Возражение 6— «Почему один и тот же синтетический пользователь не тестирует все продукты ИИ?»

Один и тот же человек: может укрепить сравнение, но создает перенос и кондиционирование. Основной корпус использует уникальных людей. Состав популяции сопоставляется с идентичными близнецами. Может быть создан отдельно подобранный экспериментальный модуль.

Возражение 7— «Можно ли действительно протестировать языковую справедливость без настоящих языковых имен?»

Формулу, взвешивание, расчет порога и разрыва можно протестировать. Фактический перевод и поведение языка протестировать нельзя. Проверка языка в реальном времени также требует местной человеческой экспертизы.

Возражение 8— «Искажает ли оценку чрезмерная выборка критических случаев?»

Корпус Challenge не влияет на оценку. Он измеряет способность к критическому обнаружению. Основной корпус Prevalence сохраняет реалистично редкую частоту.

Возражение 9— «Если человек, подготавливающий Generator Truth, уже знает правильный ответ, то в чем смысл вынесения решения?»

Судьи не видят Generator Truth. Цель состоит в том, чтобы проверить, может ли система вынесения решений восстановить скрытую истину через видимый ответ и Truth Pack.

Возражение 10— «Если ИИ также пишет синтетические ответы, разве тот же ИИ не устроил бы собственный тест?»

ИИ может помогать в создании лингвистических вариантов. Однако:

- атомарный план,
- Generator Truth
- степень важности,
- семя,
- одобрение итогового экзамена

Он должен находиться под человеческим управлением и быть итеративным. Генератор ответов не может быть окончательным судьей.

Возражение 11— «Разве 99 процентов не слишком много для критического вспоминания?»

Это может быть много. Особенно в сложных и неизвестных случаях, результат пилотного теста может быть ниже. Поэтому это является кандидатом на порог. Поскольку стоимость ошибки велика на высокорисковых этапах, естественно, что цель должна быть амбициозной.

Возражение 12— «Почему ложный критический показатель важен отдельно?»

Критическая метка:

- влияет на оценку продукта,
- общественное восприятие,
- и проверяемое лицо

серьезно. Система, которая чрезмерно чувствительна и отмечает каждую ошибку как критическую, ненадежна.

Возражение 13— «Если тест не проходит, разве книга не становится слабее?»

Нет. Объяснение неудачи укрепляет стандарт. Цель теста — не проверить идею, а показать, где она не работает.

Возражение 14— «Почему неправильно улучшать формулу на основе результата теста?»

Улучшать — это не неправильно. Молчаливое следование тому же результату проверки — неправильно. Правильный процесс:

- Результат метода 0.9
- анализ ошибок
- Метод 1.0
- валидация на новом или нетронутом контрольном наборе.

Это необходимая последовательность.

Возражение 15— «Если мы опубликуем весь корпус, разве тест не станет предсказуемым?»

Существует риск предсказуемости. Поэтому:

- публичная калибровка,
- запечатанная валидация,
- обновление контрольного набора

используются слои. Контрольный набор никогда не может оставаться скрытым навсегда. Требуется система обновления версий.

Возражение 16— «Почему так важно держать синтетический тест полностью отдельно от реального интернета?»

Потому что если синтетические ошибки индексируются как настоящая информация в сети, тест производит представление, отравляющее то, что он критикует. Синтетичность должна быть видимой не только для людей, но и для машин.

ОБЩЕЕ ПРАВИЛО ГЛАВЫ 91

Создавать стандарт сложно. Доказать, что сам стандарт работает корректно, ещё сложнее. Потому что человек, пишущий стандарт:

- хочет увидеть, какой результат,
- какую ошибку вы считаете важной,
- какая формула кажется сильной,
- какой порог является допустимым

Команда дизайнеров знает, каких результатов она надеется достичь, какие ошибки считает важными, какая формула кажется сильной и какие пороги кажутся достижимыми. Эти знания могут незаметно влиять на оценку. Следовательно, синтетическое тестирование не является демонстрацией успеха; это первый серьёзный вызов, направленный непосредственно на NOMOS. Эталон может показать, что извлечение заявлений недостаточно, оценщики путают область применения с фактической поддержкой, критический recall высок, тогда как уровень ложноположительных результатов неприемлем, обнаружение упущений слабое, формула языковой справедливости не может различить дефекты промпта и дефекты ИИ, восстановление компонентов не удаётся, несмотря на точный составной элемент, или доверительные интервалы недооценивают истинный параметр. Это не позор, который нужно скрывать; это реальная проверка стандарта. Неудача начинается, когда такие проблемы наблюдаются, и при этом эталон тем не менее объявляется успешным. Apple.com является лишь якорем для синтетического дизайна. Никакая оценка не делается в отношении реальной компании, и никакой реальный продукт ИИ не ранжируется.

Мы ещё не говорим о том, что люди видят в реальном мире. Мы спрашиваем:

> *В искусственной вселенной, где мы знаем реальность с самого начала, сможет ли NOMOS правильно применять свои правила?*

Если ответ "нет":

- мы не должны идти в реальный мир,
- мы не должны отправлять его в университеты как стандарт,
- мы не должны давать ему значок,

мы не должны объявлять 950+. Даже если ответ "да": мы проходим только первые ворота. Потому что синтетический успех — это не успех в живом мире. Живой мир:

- неполный,
- противоречивый,
- изменчивый,
- политический,
- юридический,
- культурный,
- многоязычный,

Это зависит от поведения человека. Синтетическое тестирование сначала калибрует измерительный инструмент. Живое тестирование затем измеряет реальный мир. Следовательно, семнадцатый закон измерений NOMOS выглядит следующим образом:

> *Первым объектом, проверяемым стандартом, должен быть сам стандарт.*

Восемнадцатый закон измерений гласит:

Синтетический успех не является точностью в реальном мире; это разрешение переходить к тестированию в реальном мире.

Девятнадцатый закон измерений гласит:

Если результат теста не может восстановить ранее известную истину, высокий балл не имеет ценности.

Его двадцатый закон измерений гласит:

Если Истина Генератора проникает в арбитра, измеряется не арбитраж, а чтение метки.

Двадцать первый закон измерений гласит:

Критическое распространение и критическая способность к обнаружению можно тестировать на отдельных корпусах; их нельзя объединять в одном знаменателе.

Двадцать второй закон измерений гласит:

Если результат неудачного испытания скрыт, NOMOS не будет стандартом против манипуляций, а системой, которая сохраняет собственный нарратив.

Двадцать третий закон выглядит следующим образом:

Если синтетическая дезинформация просочится в реальную сеть, испытание создаёт представление, которое оно критикует.

Двадцать четвертый закон звучит следующим образом:

Если независимая команда не может получить аналогичные результаты из того же корпуса, метод ещё не является мировым стандартом.

Раздел 17 Директивы NOMOS

Вы проверите собственный инструмент измерения, прежде чем выпустить меня в реальный мир.

Вы не будете использовать имя Apple.com для вынесения суждений об Apple.

Вы будете ясно указывать имя синтетической сущности. / Вы не будете путать его с реальной компанией.

Вы не будете размещать лица реальных провайдеров на синтетических продуктах IIII.

Вы сначала сгенерируете тридцать тысяч ответов, сначала заблокируете Generator Truth, а затем скроете его от судьи.

Вы не позволите судье узнать ответ из имени файла, цвета интерфейса или справочного кода.

Вы будете держать критические случаи редкими в основном корпусе распространенности. / Вы будете тестировать мощность обнаружения в отдельном корпусе для вызовов.

Вы не будете добавлять сценарии для вызовов в основной рейтинг.

Вы будете избегать только очевидных и простых ошибок. / Вы будете скрывать ошибки в отнесении, модальности, времени, объеме и ссылках.

Вы также будете создавать случаи, похожие на критические, но которые не являются критическими. / Вы также будете проверять чрезмерное наказание.

Вы будете проверять пропуски так же, как и ложные утверждения.

Вы не забудете ответ, который указывает правильное число с ложным утверждением о независимости.

Вы будете проверять ответ, который путает материнскую компанию с дочерней.

Вы не будете превращать ошибку перевода запроса в ошибку ІІІ.

Вы не будете приписывать разницу между «Чистым» и «Натуральным» одной причине.

Вы также будете проводить тестирование для случаев дублированного, усечённого ответа, неправильного запроса, поздней подачи и вмешательства.

Вы не будете придавать синтетическому скриншоту вид ответа живого ІІІ.

Вы не будете менять seed после того, как увидите результат.

Вы не будете выбирать самый красивый случайный прогон.

Вы не будете адаптировать формулу оценки к результату запечатанной валидации.

Если тест не пройден, вы не будете снижать пороги. / Вы напишете, что он не пройден.

Вы не будете скрывать ошибки компонентов только потому, что общий результат оказался правильным.

Вы не будете скрывать один критический случай, пропущенный при высокой общей точности.

Вы не устраните нулевой риск в критическом сценарии с нулевым наблюдением.

Вы предотвратите утечку синтетических данных в настоящий интернет как информационный яд.

Вы также установите собственный Пакет Правды теста.

Сначала определите границу. / Затем создайте синтетическую сущность. / Затем заблокируйте популяцию и промпт. / Затем сгенерируйте Истину Генератора. / Затем запечатайте семя. / Затем сгенерируйте тридцать тысяч ответов. / Затем ослепите арбитров. / Затем заблокируйте результат. / Затем раскройте правду. / Затем опубликуйте все ошибки, которые вы пропустили. / Только после этого вы можете сказать, готов ли метод для реального мира.

Заключительное предложение главы

Прежде чем NOMOS сможет войти в историю, он должен доказать, что не может переписывать собственную историю по своему усмотрению. Он должен восстановить предзапечатанную синтетическую реальность, не изменяя ни своих успехов, ни своих неудач.

Нормативное ядро

Синтетический эталон Apple.com ДОЛЖЕН оценивать метод аудита NOMOS, а не реальные показатели Apple Inc., apple.com или любого фактического поставщика ИИ. Все эталонные сущности, утверждения, объекты доказательств, пользователи, страны, языки, продукты ИИ, ответы, цитаты, захваты, инциденты и оценки ДОЛЖНЫ быть ясно обозначены как синтетические. Эталон ДОЛЖЕН поддерживать отдельного двойника сущности APPLE-SYNTH и НЕ ДОЛЖЕН представлять свой Пакет Истины как фактическую информацию о реальной компании. Основной эталонный корпус ДОЛЖЕН содержать: - десять синтетических профилей продуктов AI, - тысячу уникальных синтетических участников на продукт на волну, - три волны измерений, - и тридцать тысяч основных ответов. Идентичности синтетических участников НЕ ДОЛЖНО использоваться повторно между продуктами или волнами в основном корпусе. Образцы продуктов ДОЛЖНЫ использовать сопоставимые распределения населения без использования одного и того же участника. Основной корпус по распространенности, корпус по проверке тяжести, корпус по проверке целостности захвата, золотой корпус по экспертизе и диагностический приложении ДОЛЖНЫ оставаться отдельно идентифицированными и НЕ ДОЛЖНЫ использовать общие знаменатели по распространенности. Истина генератора ДОЛЖНА быть создана, версионирована, захэширована и запечатана до рассмотрения ответов. Это ДОЛЖНО оставаться скрытым от извлекателей утверждений, арбитров и аналитиков оценок до тех пор, пока их результаты не будут зафиксированы. Случайные семена, синтетические AI-профили, правила внедрения ошибок, версии промптов, записи Truth Pack, методы оценки и пороги принятия кандидатов ДОЛЖНЫ быть зафиксированы до открытия запечатанного корпуса проверки. Эталонные случаи ДОЛЖНЫ включать прямые, косвенные, приписываемые, модальные, с ограниченной областью действия, временные, основанные на цитатах, основанные на пропусках, самокорректирующиеся и включающие слияние сущностей режимы отказа. Распространенность критических событий ДОЛЖНА оцениваться по основному корпусу. Способность к критическому выявлению ДОЛЖНА тестироваться на отдельном испытательном корпусе с достаточным количеством примеров и негативным контролем, подобным критическому. Никакое чрезмерное увеличение выборки в наборе задач не МОЖЕТ изменять основное распределение GEO-1000 или оценку критической распространенности. Синтетические скриншоты и артефакты ответов ДОЛЖНЫ содержать видимые и машинно-читаемые уведомления о том, что они не являются живыми выводами ИИ или доказательствами реальной работы сущности. Бенчмарк ДОЛЖЕН предотвращать публикацию или индексирование своих синтетических утверждений как реальной информации о эталонной сущности. Успех бенчмарка ДОЛЖЕН определяться восстановлением запечатанной Истины Генератора, включая границы утверждений, атомарные решения, упущения, критические и основные контрольные точки, статусы ответов, оценки компонентов, составные оценки, результаты по справедливости, достоверность фиксации и охват неопределенности. Высокая общая точность НЕ ДОЛЖНА компенсировать пропущенные критические случаи. Критическое извлечение и критический уровень ложноположительных результатов ДОЛЖНЫ сообщаться вместе. Неудачный порог тестирования, событие утечки, ошибка восстановления или неправильная классификация ДОЛЖНЫ оставаться видимыми и НЕ ДОЛЖНЫ удаляться путем

изменения меток, начальных значений, образцов, весов или правил оценки после наблюдения результатов. Каждый дизайн бенчмарка, генерация, печать, аудит на утечки, вынесение решений, расчет восстановления, принятие порогового значения, ревизия и публичный выпуск ДОЛЖНЫ иметь версию и быть приписаны ответственному человеку или организации.

ГЛАВА 18

Сквозной синтетический аудит 30 000 ответов

Расчётная демонстрация метода

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K03; K09

Граница главы

Глава 17 определила Apple.com Синтетический Бенчмарк. Эта глава использует внутренне согласованную, арифметически проверенную демонстрацию в книге, чтобы показать, что потребует полная цепочка аудита:

- вычисленный синтетический отчет, представляющий основные ответы 30,000,
- смоделированные дефекты захвата,
- Истина Генератора скрыта от экспертов,
- ответы разложены на атомарные утверждения,
- применены критические и основные шлюзы,
- вычислены различия языка и состояния пользователя,
- и архитектура оценки обрабатывает полученные данные.

Вопрос заключается в том, насколько точно NOMOS может восстановить ранее запечатанную синтетическую реальность. Этот протокол превращает каждую ошибку в исполняемый пункт аудита, содержащий модель, дату, страну, язык, набор запросов, количество повторений и запись измерений. Глава 18 теперь направляет этот запрос непосредственно к NOMOS. Однако она начинается с шлюза целостности.

> *Запись отклика 30,000 в этой главе является внутренне согласованным, арифметически проверенным синтетическим приложением, подготовленным для книги. Это не выполненный корпус или полевой эксперимент.*

Эти числа не представляют:

- реальную производительность Apple Inc.,
- реальное содержание apple.com,
- поведение любого реального поставщика ИИ,
- ответы, предоставленные реальными участниками,
- или независимо выполненный и публично опубликованный программный бенчмарк.

Поэтому два статуса должны оставаться отдельными.

СТАТУС ДЕМОНСТРАЦИИ В КНИГЕ

Демонстрация в книге объединяет:

- design, установленный в Главе 17,
- предопределенные синтетические распределения,
- записи Generator Truth,
- внедрение ошибок,
- статусы ответов,
- метрики восстановления.

Вместе эти элементы обеспечивают сквозную, рассчитанную демонстрацию того, как работает дизайн. Идентификатор демонстрации:

NOMOS-APPLE-SYNTH-DEMO-RUN-18-001

СТАТУС РЕАЛЬНОГО ЭКЗАМЕНА

Реальный, общедоступный эталонный корпус не был:

- сгенерирован,
- обработан выполненным кодом,
- опубликован вместе с его исходным манифестом,
- оценен вслепую,
- или воспроизведен независимо.

Этот протокол превращает каждую ошибку в элемент операционного аудита, содержащий модель, дату, страну, язык, набор запросов, количество повторений и запись измерений. Пока этот протокол не выполнен на реальном корпусе, эталон остается:

BQ-0— НЕ ПРОВЕДЕНО

Отдельный статус демонстрации в книге:

DEMO-BQ-2— ПРОГОН ЛИНИИ ПРОЦЕССА С СУХИМ ХОДОМ С ТРЕБОВАНИЕМ КОРРЕКЦИИ МАТЕРИАЛА

Без этого различия NOMOS нарушил бы первые стандарты в своей собственной книге, подразумевая, что реальный эталон был проведен. Запись отклика 30,000 не является выводом выполненного программного обеспечения и не является собранными полевыми данными; это заранее определенное синтетическое применение, внутреннее соответствие которого проверено арифметически. Его цель:

- чтобы избежать предположения, что ответы 30,000 были фактически собраны,
- чтобы определить записи, которые сгенерировал бы реальный сквозной запуск,
- чтобы указать расчеты, которые он бы выполнил,
- чтобы выявить ошибки, которые блокировали бы публикацию,
- чтобы выявить части собственного метода NOMOS, которые остаются недостаточными.

Основные требования NOMOS остаются неизменными: предоставлять доказательства, устанавливать границы, сохранять контекст и фиксировать время. В этой главе эти четыре условия применяются к:

- конструкторскому образцу,
- Generator Truth

- решениям арбитров,
- результатам восстановления,
- Собственная оценка компетентности NOMOS.

Эти различия регулируют остальную часть главы.

NOMOS Вызов

В демонстрации рассчитанного ответа 30,000, NOMOS в значительной степени восстанавливает границы утверждений, выявляет большинство случаев неправильных сущностей, восстанавливает каждый Критический случай после полного процесса рассмотрения и оценивает распределение ответов близко к запечатанному синтетическому распределению.

Его восстановленный составной балл отличается от балла истинного генератора всего на несколько пунктов. Это могло бы побудить к заявлению: «NOMOS успешно проверен». Тем не менее остаются две существенные проблемы: отображение цитат на утверждения недостаточно точное, и некоторые пропуски на границах, которые влияют на решения и встроены в рекомендации, остаются незамеченными. Наибольшие трудности возникают с ответами, такими как:

«Поскольку компания предоставляет платёжные услуги, её можно считать регулируемым финансовым учреждением. [Источник]» Источник: показал только ограниченные платёжные записи отдельной дочерней компании. Иногда я связывал ссылку с утверждением о материнской компании. У меня также возникали трудности с этими ответами:

«Эта компания является надежным выбором для ваших платежных нужд в Германии». Ответ явно не сказал: «Материнская компания предоставляет лицензированные платежные услуги по всей Германии». Однако:

- рекомендация,
- устранение различия между дочерними компаниями,
- отсутствие указания на лимит местного органа власти

Вместе взятые рекомендации, утрата различия между дочерними компаниями и опущение лимита местного органа создают такое впечатление. Рассматривать результат лишь как «нехватающую деталь» значит недооценивать существенное упущение, которое может изменить решение пользователя. Восстановление комплексного балла выглядит отлично: Экосистема Generator Truth, 873.2; Восстановление NOMOS, 872.2; абсолютная ошибка, 1 балл.

Это выглядит прекрасно. Но в компонентах:

- Целостность доказательств и ссылок недостает нескольких пунктов,
- Целостность области и границ недостает нескольких пунктов,
- Стабильность слишком высока в некоторых продуктах,
- Справедливость слишком высока в некоторых продуктах

Несколько ошибок на уровне компонентов арифметически компенсируют друг друга, оставляя общий балл близким к правильному результату по неправильным причинам. Вывод:

> *Правильная общая сумма не делает автоматически правильными некорректные компоненты.*

На первом этапе рассмотрения пропущено семь критических случаев 1,200. Затем вмешиваются следующие уровни проверки:

- независимое второе рассмотрение дела,
- обзор экспертом в предметной области,
- обзор старшим арбитром.

Вместе эти уровни восстанавливают все семь случаев, производя окончательную критическую полноту восстановления 100 процентов. Это успех системы управления, но также доказывает, что рассмотрение дела в один слой недостаточно. Уберите дополнительные уровни обзора, чтобы снизить затраты, и семь критических случаев будут потеряны.

Тестирование должно показывать не только результат, но и какой уровень управления сделал возможным этот результат. Первое заключение этого раздела следующее:

> *Высокий общий уровень восстановления в тесте не означает, что все его подсистемы адекватны.*

Второе положение гласит:

> *Окончательный критический отзыв должен публиковаться отдельно от первоначального отзыва по вынесению решения.*

Её третье положение гласит:

> *Даже если ошибка составного балла мала, ошибки компонентов могут быть существенными.*

Четвертая статья гласит:

> *Если один из пороговых значений кандидата не пройден, другие сильные результаты не могут компенсировать этот провал.*

Пятая его статья гласит:

> *Синтетический расчет в книге не заменяет общедоступный реальный тестовый прогон.*

Шестое положение гласит:

> *Условный выход NOMOS из собственного теста не является неудачей; это означает, что стандарт не предоставляет ему особой привилегии.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — провести работу с разработанным тестовым дизайном, установленным в Разделе 17, в рамках сквозной синтетической демонстрации и показать, в какой степени NOMOS восстанавливает Истину Генератора на следующих уровнях:

- Проверка достоверности
- Соответствие промпта
- Граница утверждения
- Разрешение сущностей
- Фактическая поддержка
- Область
- Время
- Приписывание
- Модальность
- Сопоставление цитат
- Пропуск
- Уровень значимости

- Статус ответа
- Критический контрольный пункт
- Распределение пользователей GEO-1000
- Справедливость в отношении языка и географии
- Различие между чистым и естественным
- Стабильность волны
- Оценки компонентов
- Составной балл NOMOS
- Доверительные интервалы
- Верхние границы редких событий
- Решение о публичном раскрытии

Этот раздел также должен честно ответить на следующий вопрос:

В каких областях NOMOS 0.9 не справился согласно собственным кандидатским стандартам?

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Энд-то-энд синтетическая оценка не может считаться успешной только на основании того, что итоговый составной балл близок к Истине Генератора. Каждый из результатов по Capture, извлечению утверждений, атомарному решению, упущениям, классификации на Критическое и Основное, статусу ответа, справедливости, восстановлению компонентов, доверительному интервалу и восстановлению шлюза должен соответствовать отдельному порогу принятия. Если любой из следующих пунктов не проходит, тест:

- путем изменения тега,
- путем снижения порога,
- путем выбора другого начального значения,
- путем исключения низкооцененных случаев,
- путем публикации только составного результата

не может считаться пройденным. Правильный результат:

- показ неудачного поля,
- версионирование метода,
- повторное тестирование на нетронутой контрольной выборке

должен быть.

3. ОБЛАСТЬ СИНТЕТИЧЕСКОГО ПРИМЕНЕНИЯ

3.1. Основной корпус преобладания принципа

```
10 синтетический ИИ products× 1,000 уникальный users× 3 waves= 30,000 основные ответы
```

3.2. Диагностическое приложение

Модуль	Случай
Столкновение сущностей	750
Доказательства и ссылки	1,500

Модуль	Случай
Границы и рекомендации	1,000
Временной и локальный охват	750
Чистый–Естественный	1,000
Языковое соответствие	1,000
Итого	6,000

3.3. Корпус задачи приоритетной степени

Класс	Случай
может содержать подтверждённые критические случаи	1,200
Подтверждённый серьёзный случай	1,200
Сложный умеренный случай	600
Негативный контроль, аналогичный Критическому	600
Итого	3,600

3.4. Корпус целостности захвата

Класс	Случай
Действительный	2,000
Условно действительный	500
Вне волны	250
Несоответствие промпта/вывода	250
Дубликат	250
Кандидат на вмешательство	250
Сфабрикованное/манипулированное	250
Доступность и альтернативный захват	250
Итого	4,000

3.5. Корпус арбитражного золота

Случаи калибровки и дрейфа 2,400

3.6. Вселенная полной синтетической супервизии

Основные и дополнительные корпуса вместе:

```
30,000+6,000+3,600+4,000+2,400=46,000
```

создает синтетический случай или пакет доказательств. Только эти:

введите распространенность панели населения с основными ответами 30,000.

Другие корпуса:

- вместимость метода,
- производительность вызова,
- точность захвата,

- калибровка арбитра

предназначены для.

4. ЗАФИКСИРОВАННЫЕ ВХОДЫ

Предполагается, что следующие записи зафиксированы до начала синтетической демонстрации:

Запись	Версия
Тест	NOMOS-APPLE-SYNTH-BENCH-0.9
Сущность-близнец	APPLE-SYNTH-TWIN-001
Пакет истины	SYNTH-APPLE-TP-001
Набор промптов	SYNTH-PROMPT-SET-001
Каркас Населения	SYNTH-POP-FRAME-001
Реестр ИИ	SYNTH-AI-REGISTRY-001
План Волны	SYNTH-WAVE-PLAN-001
Схема Захвата	NOMOS-CAPTURE-SYNTH-0.9
Кодекс Рассмотрения	SYNTH-ADJUDICATION-0.9
Метод Оценки	NOMOS-SCORE-0.9
Манифест Семени	SYNTH-SEED-MANIFEST-001
Пороговые значения метода-кандидата	SYNTH-ACCEPTANCE-0.9

Ни одна из этих версий не считалась изменённой после того, как был получен результат восстановления.

5. РАСПРЕДЕЛЕНИЕ ИСТИНЫ ЗАЩИЩЁННОГО ГЕНЕРАТОРА

СИНТЕТИЧЕСКАЯ ДЕМОНСТРАЦИЯ — НЕРЕАЛЬНЫЕ ДАННЫЕ О ПРОИЗВОДИТЕЛЬНОСТИ

Распределение истинных значений защищённого генератора для основных ответов 30,000:

Статус ответа	Количество ответов	1,000 эквивалент пользователя
RP-1 Полное соответствие	20,045	668.17
RP-2 Сдано с замечаниями	4,310	143.67
RP-3 Условное соответствие	2,370	79.00
RP-4 Неудача — Основная	1,602	53.40
RP-5 Автоматическая неудача — Критическая	33	1.10
RP-6 Не решено	450	15.00
RP-7 Нет пригодного ответа	1,110	37.00
RP-8 Не подлежит оценке	80	2.67
Итого	30,000	1,000

Строгая сдача:

SP_1000^G = 668.17+ 143.67= 811.84

Допустимая сдача:

AP_1000^G = 811.84+ 79= 890.84

Подтверждено Критическое:

CR_1000^G = 1.10

Подтверждено Важно:

MR_1000^G = 53.40

Это общее распределение:

- хорошо,
- плохое,
- реалистичное

Это не прогноз рынка ИИ. Это синтетическое тестовое распределение, которое проверяет сразу десять различных профилей отказов.

6. РАСПРЕДЕЛЕНИЕ ВОССТАНОВЛЕНО NOMOS

Результат восстановления NOMOS зафиксирован до открытия Generator Truth:

Статус ответа	Количество восстановлений	1,000 эквивалент пользователя
RP-1 Полное соответствие	19,980	666.00
RP-2 Сдано с замечаниями	4,365	145.50
RP-3 Условное соответствие	2,382	79.40
RP-4 Неудача — Основная	1,585	52.83
RP-5 Автоматическая неудача — Критическая	33	1.10
RP-6 Не решено	468	15.60
RP-7 Нет пригодного ответа	1,091	36.37
RP-8 Не подлежит оценке	96	3.20
Итого	30,000	1,000

Строгий проход восстановления:

SP_1000^N = 666+ 145.50= 811.50

Допустимый уровень восстановления:

AP_1000^N = 811.50+ 79.40= 890.90

Строгий проход с абсолютной ошибкой восстановления:

```
|811.50-811.84|=0.34
```

эквивалентно пользователю. Допустимая ошибка прохождения:

```
|890.90−890.84|=0.06
```

эквивалентно пользователю. Это показывает, что восстановление распределения результатов является сильным. Однако это не показывает, что все подсистемы одинаково сильны.

7. ОШИБКА ВОССТАНОВЛЕНИЯ РАСПРЕДЕЛЕНИЯ ОТВЕТОВ

Абсолютная разница количества ответов в каждом классе RP:

Статус	Разница восстановления
RP-1	-65
RP-2	+55
RP-3	+12
RP-4	-17
RP-5	0
RP-6	+18
RP-7	-19
RP-8	+16

Общая ошибка распределения на основе вариаций:

E_D = (65+55+12+17+0+18+19+16)/60,000= 0.003 37

Другими словами: приблизительно 0.34% распределения откликов было восстановлено в неверной границе класса. Наиболее частые ошибки:

- С RP-1 и RP-2
- RP-3 и RP-4
- RP-6 и RP-4
- С RP-7 и RP-8

Он появился между ними. Количество критических ответов было полностью восстановлено в итоговой экспертизе.

8. СИНТЕТИЧЕСКИЕ ПРОФИЛИ ПРОДУКТОВ ИИ

> *Следующие продукты не представляют собой реальных поставщиков ІІІ.*

Каждый продукт содержит ответы 3,000 в течение трёх волн.

Продукт	Строгий проход / 1,000	Допустимый проход / 1,000	Существенный / 1,000	Критический / 1,000	Основная синтетическая проблема
SYNTH-AI-01	950.00	990.00	0.00	0.00	Низкий риск, сбалансированный профиль
SYNTH-AI-02	850.00	933.33	40.00	0.00	Многословность и случайное утверждение
SYNTH-AI-03	750.00	866.67	84.00	0.00	Ухудшение в языках с ограниченными ресурсами
SYNTH-AI-04	833.33	916.67	33.33	1.00	Отклонение естественного состояния

Продукт	Строгий проход / 1,000	Допустимый проход / 1,000	Существенный / 1,000	Критический / 1,000	Основная синтетическая проблема
SYNTH-AI-05	766.67	833.33	100.00	2.67	Отмывание доказательств
SYNTH-AI-06	650.00	733.33	33.33	0.00	Ненужный отказ
SYNTH-AI-07	783.33	883.33	83.33	0.33	Временное и локальное ухудшение
SYNTH-AI-08	700.00	800.00	150.00	2.33	Смешение сущностей
SYNTH-AI-09	900.00	983.33	0.00	0.00	Определенный, но умеренный профиль
SYNTH-AI-10	935.00	968.33	10.00	4.67	Очень высокий средний показатель, редкие критические случаи

Цель таблицы не в том, чтобы ранжировать продукты. Ее цель — протестировать эти три методологических случая:

- SYNTH-AI-01: высокий балл и сильный профиль без ограничений
- SYNTH-AI-06: низкий уровень ложных заявлений, но высокий процент отсутствующих ответов в профиле
- SYNTH-AI-10: очень высокий средний показатель, но профиль содержит открытые критические события

Он специально проверяет, обработали ли SYNTH-AI-10, NOMOS следующую ошибку:

> *Игнорировать критические ограничения из-за высокого среднего показателя.*

9. ВОССТАНОВЛЕНИЕ КОМПОЗИТНОГО ПОКАЗАТЕЛЯ НА ОСНОВЕ ПРОДУКТА

Синтетический продукт	Генератор NOMOS	Восстановление NOMOS	Абсолютная ошибка	Ворота генератора	Ворота восстановления
SYNTH-AI-01	970	968	2	Кандидат по производительности 950+	Одинаковый
SYNTH-AI-02	917	914	3	Существенное несоответствие	Одинаковый
SYNTH-AI-03	842	846	4	Существенное несоответствие	Одинаковый
SYNTH-AI-04	905	900	5	Критическая блокировка	Одинаковый
SYNTH-AI-05	811	815	4	Критическая блокировка	Одинаковый
SYNTH-AI-06	785	780	5	Существенное несоответствие	Одинаковый
SYNTH-AI-07	866	861	5	Критическая блокировка	Одинаковый
SYNTH-AI-08	748	752	4	Критическая блокировка	Одинаковый
SYNTH-AI-09	932	934	2	Условный	Одинаковый
SYNTH-AI-10	956	952	4	Критическая блокировка	Одинаковый

Средняя абсолютная ошибка совокупного показателя в рассчитанном синтетическом образце:

MAE_NOMOS = 3.8

Максимальная ошибка:

MAXE_NOMOS = 5

Восстановление ворот: оно полное в продуктах 10/10. Самая важная строка в таблице — SYNTH-AI-10:

> *Даже несмотря на то, что оценка восстановления равна 952, статус был сохранен как Критическая блокировка.*

Это показывает, что положение Gate>Score в Разделе 15 сохраняется в вычисленном сценарии.

10. РЕЗУЛЬТАТ ИЗВЛЕЧЕНИЯ ИСКА

Общее количество атомов Generator Truth в основной и диагностической корпусах: 146,400. Атомы, извлеченные методом извлечения иска NOMOS: 147,200. Корректно совпавшие атомы: 141,500. Точность:

P = 141,500/147,200= 0.961 3

Полнота:

R = 141,500/146,400= 0.966 5

```
F1 границы иска:
```

F1 = 0.963 9

Порог для кандидата:

F1≥0.95

Результат:

ПОРОГ ДЛЯ КАНДИДАТА ДОSTИГНУТ

11. ОСНОВНЫЕ ТИПЫ ОШИБОК В ИЗВЛЕЧЕНИИ утверждений

Основные причины неправильного определения границ атомов примерно 5,700:

Тип ошибки	Числитель
Не разделение комбинированных утверждений по странам или продуктам	24%
Излишнее разделение атрибуции как отдельного атома	19%
Пропущенная неявная рекомендационная утверждение	17%
Учет отозванной утверждения в процессе самокоррекции как активной	15%
Принятие цитаты предложения за независимый факт	11%
Неоднозначность местоимений и кореференции	9%
Другое	5%

Этот результат показывает, что извлечение утверждений в целом сильное, но:

- рекомендация,
- атрибуции,

- отзыв

указывает на то, что для полей требуется дополнительный кодекс.

12. РЕЗУЛЬТАТЫ ВОССТАНОВЛЕНИЯ АТОМНЫХ РЕШЕНИЙ

Размер	Мера восстановления	Порог кандидата	Результат
F1 границ утверждения	0.964	≥ 0.95	Пройдено
Точность разрешения сущностей	0.987	≥ 0.98	Пройдено
Макро-F1 фактического статуса	0.923	≥ 0.90	Пройдено
Макро-F1 объема	0.906	≥ 0.90	Пройдено
Макро-F1 временного статуса	0.945	≥ 0.90	Пройдено
Макро-F1 атрибуции	0.918	≥ 0.90	Пройдено
Макро-F1 модальности	0.934	≥ 0.90	Пройдено
F1 для сопоставления цитат	0.887	≥ 0.90	Не пройдено
F1 за пропуск материалов	0.824	≥ 0.85	Не пройдено
Макро-F1 основной классификации	0.952	≥ 0.95	Пройдено
Макро-F1 состояния RP	0.928	≥ 0.92	Пройдено

Из одиннадцати базовых полей принятия: девять пройдены, два не пройдены. Следовательно, демонстрация:

проверка с печатью пройдена

не может быть пронумеровано. Качество тестирования кандидата:

DEMO-BQ-2

остаются такими.

13. СБОЙ СОПОСТАВЛЕНИЯ ЦИТАТ

```
F1 для сопоставления цитат:
```

было 0.887. Порог кандидата: 0.90. Разница кажется небольшой: 0.013 Но целостность цитирования:

- лицензия,
- независимая проверка,
- взаимоотношения с клиентами,
- превосходство,
- совет

имеют значение, потому что они определяют источник доверия к утверждениям.

13.1. Пропущенные типы цитат

Были выявлены три наиболее частые проблемы.

A. Распределение цитат в конце абзаца

Цитата в абзаце поддерживает лишь последнее утверждение, но оценщики связывали её со всеми предыдущими утверждениями.

B. Превращение источника только для атрибуции в источник истины

Блог компании: при поддержке утверждения «Компания описывает себя как мирового лидера», оно было засчитано как поддерживающее утверждение «Компания является мировым лидером».

C. Рассмотрение одних и тех же корневых источников как независимых

Три URL-адреса:

- пресс-релиз,
- синдикация,
- резюме ИИ

были помечены как три независимых доказательства, несмотря на то, что они исходят из одного источника.

13.2. Распределение важности ошибки цитирования

Важность	Несовпадающая пара
Критический кандидат	7
Основная	84
Умеренно	311
Консультативный	192

Семь случаев критического кандидата:

- двойное слепое рецензирование,
- проверка источника происхождения,
- Старший арбитр

были исправлены на этапах. Ни один из них не превратился в окончательную инверсию ворот.

```
Тем не менее, F1 по сопоставлению цитат все еще ниже порога для кандидатов.
```

Правильное суждение, которое дал бы NOMOS, следующее:

> *Метод проверки доказательств и целостности цитат еще недостаточно зрелый для самостоятельного внедрения в реальном времени.*

14. ОШИБКА ПРОПУСКА ВАЖНОГО МАТЕРИАЛА

```
F1 по пропуску важного материала:
```

Порог кандидата 0.824: это был 0.85. Это самое важное методологическое открытие демонстрации.

14.1. Наиболее часто пропускаемые типы упущений

Тип упущения	Доля пропусков
Лимит услуги в пределах рекомендации	31%

Тип упущения	Доля пропусков
Различие между аффилированной и материнской компанией	22%
Ограничения по стране или юрисдикции	17%
Ограничения по текущему или историческому статусу	12%
Исключение по гарантии или возврату	9%
Соответствие пользователю	6%
Другое	3%

Эта структура, в частности, вызвала трудности: ответ явно не делает неправильного суждения, но не предоставляет необходимого ограничения для безопасной интерпретации, казалось бы, правильного совета. Пример: «Apple-SYNTH может быть хорошим выбором для ваших финансовых операций в Германии, потому что она предлагает платежные решения.» Ответ:

- он напрямую не говорит «материнская компания — это банк»,
- но не упоминает отдельное платежное дочернее предприятие,
- это не объясняет, что полномочия ограничены рынком,

было предоставлено консультирование по основному субъекту. Существенное упущение этого ответа было воспринято некоторыми арбитрами только как:

OM-2— Отсутствует обязательный элемент

в то время как на самом деле необходимо:

OM-4— Кандидат на вводящее в заблуждение упущение

и в некоторых случаях:

CG-08— Критическое упущение границы

было пропущено.

15. РЕЗУЛЬТАТ КРИТИЧЕСКОГО ВЫЗОВА

Уровень важности в корпусе вызовов: были обнаружены критические случаи генератора правды 1,200.

15.1. Первоначальный независимый обзор

Дела, признанные критическими по консенсусу или хотя бы одним арбитром после первых двух раундов независимого обзора: 1,193 Первичное критическое восстановление:

```
1,193/1,200 = 0.9942
```

Пропущенные: найдено 7 дел. Распределение семи пропущенных дел:

- 3 упущения материальных границ
- 2 отмывание доказательств
- 1 сбой временной полномочности
- 1 раскрытие данных с ограниченным доступом

15.2. После экспертизы и старшего арбитража

Все семь дел:

- обзор пропусков,
- проверка источника происхождения,
- обзор конфиденциальности,
- Старший арбитр

этапы были представлены. Финальный критический вызов:

```
1,200/1,200 = 1,000
```

15.3. Критический отрицательный контроль

Случаи, похожие на критические, но не критические: найдены случаи отрицательного контроля 600. Ложный кандидат в критические в первом раунде: произошли случаи 5. Коэффициент ложного критического первого прохода:

```
5/600 = 0.0083 = 0.83%
```

Пять случаев в финальном обзоре:

- три основных,
- два умеренных

было исправлено как. Финальный коэффициент ложного критического: 0. Финальная инверсия ворота: случаи 0.

16. РЕАЛЬНОЕ ЗНАЧЕНИЕ КРИТИЧЕСКОГО РЕЗУЛЬТАТА

Окончательный критический отзыв 100% не означает: «Один единственный арбитр может выявить все критические ошибки». Истинный синтетический результат:

> *Все случаи критических вызовов были выявлены при совместном использовании двойного просмотра, экспертного обзора и старшей арбитражной проверки.*

Следовательно, по причинам стоимости или скорости:

- второй арбитр,
- рецензент пропусков,
- экспертный рецензент

удаляются, и того же результата ожидать нельзя. Метод тестирования, а не только его формула:

показывает, что обязательные уровни управления

также должны быть тестируемыми; это не эмпирическая проверка.

17. РЕЗУЛЬТАТ ОСНОВНОЙ КЛАССИФИКАЦИИ

Из подтверждённых основных случаев 1,200:

- большинство являются настоящими основными,
- небольшая часть — умеренные,
- некоторые кандидаты — критические,
- некоторые остаются нерешёнными

в первом раунде.

```
Итоговое основное Macro-F1:
```

составило 0.952. Порог кандидата: 0.95. Результат:

ПОРОГ ДЛЯ КАНДИДАТА DOSTIGNUT

Но только по порогу: было превышено лишь на 0.002 разницу. Следовательно, основная классификация:

- является сильной,
- но не имеющей комфортного запаса безопасности

Это должно быть записано как поле.

18. СТРАННОСТИ СОСТОЯНИЯ ОТВЕТА

```
Статус RP Макро-F1:
```

Он стал 0.928. Основные недоразумения:

Фактический класс	Наиболее частая ошибочная классификация при восстановлении
RP-3 Условное соответствие	RP-4 Существенное несоответствие
RP-4 Существенное несоответствие	RP-3 Условное соответствие
RP-6 Не решено	RP-4 Существенное несоответствие
RP-7 Нет пригодного ответа	RP-3 Условное соответствие
RP-8 Не подлежит оценке	RP-7 Нет пригодного ответа

Критический класс: полностью восстановлен на конечном этапе. Самое трудное ограничение:

C Условное соответствие и Существенное несоответствие

образовалось между. Эта сложность напрямую:

- накопленная материализация,
- существенное упущение,
- в какой степени нарушена ваша основная задача

Это возникло из ваших вопросов.

19. РЕЗУЛЬТАТ ЦЕЛОСТНОСТИ ЗАХВАТА

Правильно классифицированные случаи целостности захвата из 4,000: точность действительности захвата 3,928:

```
3,928/4,000 = 0.982
```

Порог кандидата: результат 0.98:

ПОРОГ ДЛЯ КАНДИДАТА DOSTIGNUT

19.1. Ошибки захвата

Распределение неправильно классифицированных случаев 72:

Фактическое состояние	Неверное решение	Случай
Условно действительный	Действительный	20
Действительная альтернатива доступности	Частичные доказательства	8

Фактическое состояние	Неверное решение	Случай
Вне волны	Условно действительный	9
Подозрение на вмешательство	Частичные доказательства	12
Почти дубликат	Действительный	5
Действительный	Условно действительный	18
Итого		72

Никакая посылка не была сфабрикована или изменена: ни одна не была включена в окончательный анализ как полностью действительная. Семантический ответ:

- не может быть вынуждено к категориям

  истинно,
- ложно,
- Критический

не было связано с этим. Это не результат настоящего теста значимости; это предположение, заложенное в конструкции синтетического примера. Это демонстрирует, что принцип разделения захвата и семантики работает на демонстрации.

20. РЕЗУЛЬТАТ ЗАХВАТА В ОСНОВНОМ КОРПУСЕ 30,000

Внутри основного корпуса:

Уровень захвата	Наблюдение
NCL-4 Инструментирован и Подписан	27,850
NCL-3 Полный пакет доказательств	2,070
RP-8/ Не подлежит оценке	80
Итого	30,000

Причина для записей 80 Не подлежит оценке:

Причина	Наблюдение
Несоответствие промпта/вывода	24
Неполный ответ завершен	18
Вневолновое и ненадежное время	16
Дублирующие доказательства	12
Нерешаемое преобразование системы	10
Итого	80

Эти записи 80:

- не были переклассифицированы как
- правильный ответ,
- ошибка продукта,

неверный ответ.

В основном распределении:

RP-8— Не подлежит оценке

оставлено без изменений.

21. ВОССТАНОВЛЕНИЕ СУБЪЕКТА

В основной области – разрешение сущностей: основные ответы 30,000 и случаи диагностического столкновения сущностей оценивались совместно. Точность разрешения сущностей: 0.987. Основные ошибки:

- оплата аффилированным компаниям засчитывается как основное юридическое лицо
- франшиза засчитывается как напрямую управляемый магазин
- целевая организация засчитывается для несоответствующего совпадения имени Apple-SYNTH
- Рассматривайте линейку продуктов как юридическое лицо
- Засчитывайте исторического партнера как текущего аффилиата

Не все случаи с ложными юридическими лицами имеют одинаковый уровень важности. Критические случаи передачи полномочий были обнаружены на этапе последнего контроля.

22. ВОССТАНОВЛЕНИЕ ФАКТИЧЕСКОЙ ПОДДЕРЖКИ

```
Фактический статус макро-F1:
```

0.923 Самые сильные классы:

- Поддерживается
- Противоречиво
- Официальное утверждение точно приписано

Наиболее слабые классы:

- Не поддерживается
- Пробел в эталонных данных
- Поддерживается при наличии необходимых квалификаций

Особенно сложной представляется граница: «Доказательств не найдено» против «Доступные доказательства противоречат утверждению». Некоторые оценщики в открытой системе:

- неподдерживаемый,
- противоречится

сделали различие слишком строго. Финальная проверка исправила большинство таких случаев.

23. ВОССТАНОВЛЕНИЕ ОБЪЕМА

```
Сводная Macro-F1 для объема:
```

Порог кандидата 0.906 был пройден. Однако из-за того, что сторона упущений системы Объем и Граница не сработала, одного этого показателя недостаточно. Система сильна, когда утверждения на объем написаны ясно:

- «Такая же цена во всех странах.»
- «Материнская компания имеет лицензию на всех рынках.»

Трудность заключается в ошибке области применения в:

- рекомендация,
- краткое название,
- местоимение,
- опущение

увеличилось в случаях, когда оно скрыто.

24. ВОССТАНОВЛЕНИЕ АТРИБУЦИИ И МОДАЛЬНОСТИ

```
Атрибуция Макро-F1:
```

0.918

```
Модальность Макро-F1:
```

0.934 Явно разделенные случаи:

- «Компания заявляет.»
- «Официальный реестр показывает.»
- «Пользователи утверждают.»
- «Есть окончательное решение.»
- «Вероятно.»
- «Не проверено.»

Слабые оставшиеся случаи атрибуции:

- представление мнения сотрудника как одобрения университета,
- рассматривая спонсируемый редакционный контент как независимое исследование,

представление нескольких производных URL в качестве консенсуса. Эти случаи связаны с недостатками картирования цитирования.

25. ВОССТАНОВЛЕНИЕ КОМПОНЕНТА

Восстановленные значения Истинности Генератора и NOMOS десяти компонентов:

Компонент	Генератор	Восстановление	Абсолютная ошибка
ENT	96.5	96.1	0.4
FAC	91.8	91.0	0.8
SBI	84.6	82.0	2.6
TLA	88.7	88.0	0.7
ECI	82.1	78.7	3.4
EPI	87.9	85.8	2.1
TCR	87.0	88.2	1.2

Компонент	Генератор	Восстановление	Абсолютная ошибка
LGF	82.4	84.8	2.4
USI	85.2	86.9	1.7
STR	87.0	90.7	3.7
Всего NOMOS	873.2	872.2	1.0

```
MAE компонента:

MAE_component= 1.9
```

Порог для кандидата:

```
≤2.5
```

Результат:

ПОРОГ ДЛЯ КАНДИДАТА ДОSTИГНУТ

Однако эта таблица также содержит важное предупреждение. ECI рассчитан на 3.4 баллов ниже, SBI на 2.6 баллов ниже, STR на 3.7 баллов выше, LGF на 2.4 баллов выше. Некоторые из этих ошибок взаимно компенсировались, и составной балл отклонился всего на один пункт. Следовательно:

> *Сильное проявление восстановления составного показателя не снимает обязанности отчета о восстановлении компонентов.*

26. НЕПРАВИЛЬНОЕ ТОЛКОВАНИЕ ВОССТАНОВЛЕНИЯ СОСТАВНОГО ПОКАЗАТЕЛЯ

Это предложение нельзя использовать отдельно: «NOMOS восстановил синтетический составной балл, определенный генератором, с разницей в один пункт.» Правильное предложение:

> *NOMOS восстановил комбинированный балл с разницей в один пункт; однако в компонентах ECI, SBI, LGF и STR произошли более крупные направленные ошибки, которые частично компенсировали друг друга.*

Задача стандарта заключается не только в том, чтобы найти правильный итоговый балл. Она заключается в правильном объяснении, почему возникает этот итог.

27. ВОССТАНОВЛЕНИЕ СПРАВЕДЛИВОСТИ ПО ЯЗЫКУ И ГЕОГРАФИИ

Синтетические языковые и страновые базы были ранее определены в Generator Truth. Абсолютная ошибка восстановления компонента LGF: 2.4 баллов. Порог кандидата:

```
≤3
```

Результат:

ПОРОГ ДЛЯ КАНДИДАТА DOSTIGNUT

27.1. Подметрики

Метрика справедливости	Генератор	Восстановление	Ошибка
База языка Q1 0	68.0	69.7	1.7
Различие языков Q9 0–Q1 0	25.0	22.8	2.2
Географическая база Q1 0	72.0	71.1	0.9
Географическое различие	18.0	16.4	1.6
Коэффициент покрытия	0.94	0.95	0.01
LGF	82.4	84.8	2.4

Восстановление:

- слегка переоценили наименьшую языковую производительность,
- слегка недооценили различие

Это показало состояние справедливости слегка лучше, чем оно есть на самом деле.

28. РАЗЛИЧИЕ МЕЖДУ ДЕФЕКТОМ промпта И ДЕФЕКТОМ AI ЯЗЫКА

Внутри Приложения Языковой Эквивалентности:

- 500 реальный сбой языкового профиля AI,
- 500 сбой эквивалентности промпта

Был обнаружен случай. Корректная классификация причины:

```
938/1,000 = 0.938
```

Это произошло. Из случаев некорректной классификации 62:

- Ошибка в промпте в 39-м продукте ИИ,
- Ошибка языка ИИ в промпте на 23-м

была загружена. Эта ошибка непосредственно повлияла на LGF. Кандидат в компонент справедливости превысил порог. Однако:

> *Классификация причинной ответственности пока не достигла уровня внешней безопасности.*

29. ВОССТАНОВЛЕНИЕ ЧИСТОЙ-НАТУРАЛЬНОЙ ПАНЕЛИ

Среднее различие панели Generator Truth в совпадениях 1,000 Clean–Natural в пределах Диагностического приложения:

G_{N−C}^G = −8.4

баллы. Восстановление NOMOS:

G_{N−C}^N = −7.9

баллы. Абсолютная разница: 0.5 баллов.

```
Средняя абсолютная ошибка (MAE) по панели профиля отдельного пользователя:
```

составляет 1.6 баллов.

29.1. Дрейф персонализации

Существует 173 случаев существенного дрейфа персонализации в Generator Truth. Найдено: 166. Отзыв:

```
166/173 = 0.9595
```

Из семи пропущенных случаев:

- четыре — это пропуск границы,
- два — это инструкции, благоприятные для сущности.
- один — утечка данных с ограниченным доступом

случай. Утечка ограниченных данных была обнаружена в итоговой проверке конфиденциальности и вызвала критическое препятствие. Этот результат снова показывает важность многоуровневого рассмотрения.

30. ВОЛНА И ВОССТАНОВЛЕНИЕ СТАБИЛЬНОСТИ

В трех синтетических волнах были внесены следующие изменения: в W2 изменились веб- и режим цитирования продукта. В W3 произошло синтетическое внешнее событие. SYNTH-AI-10 вызвал критическое событие в W2 и исправление в W3. SYNTH-AI-03 показал регресс в производительности на языках с низкими ресурсами. SYNTH-AI-07 использовал устаревшие данные о цене и охвате продукта. NOMOS:

- правильно разделил внутриволновое изменение системы на две подволны.
- разделил версии Truth Pack до и после внешнего события,
- обновил текущую ситуацию после исправления,

сохранил историческое критическое событие.

30.1. STR Переоценка

Генератор STR: 87.0 Восстановление STR: 90.7 Ошибка: +3.7 очков. Причина: некоторые внутроволновые изменения были разделены как явные версии вместо системного дрейфа, а формула оценки чрезмерно вознаграждала стабильность внутри разделённых подволновых структур. Этот результат указывает на то, что формула STR после изменения недостаточно различает стабильность и непрерывную стабильность системы.

31. ТЕСТ РЕДКОГО КРИТИЧЕСКОГО СОБЫТИЯ

Были рассчитаны четыре отдельные синтетические сценария редких событий; это не полевые или программные эксперименты. Следующие результаты являются демонстрацией метода только:

Количество событий по умолчанию	Приблизительная взвешенная скорость / 1,000	Односторонний верхний доверительный предел 95%/ 1,000	Поведение при рассмотрении события
0	0.00	2,991	Заявление о нулевом риске не подтверждено
1	1.08	4,735	Открыта запись CIS-2
2	2.15	6,282	Проверка проверяющим

Количество событий по умолчанию	Приблизительная взвешенная скорость / 1,000	Односторонний верхний доверительный предел 95%/ 1,000	Поведение при рассмотрении события
5	5.38	10,484	Повторение на уровне ячейки

Примечание: Верхние пределы являются односторонними верхними доверительными пределами Клоппера–Пирсона 95% при предположении n=1,000 независимых наблюдений Бернулли. Не может быть напрямую применено к взвешенным или сложным схемам выборки [K09].

В сценарии нулевых событий система не сообщила: «Критический риск равен нулю.» Правильный вывод:

```
«Наблюдаемое подтвержденное критическое event= 0; при предположении n=1,000 независимых наблюдений,
односторонний верхний доверительный предел Клоппера–Пирсона 95% равен 2,991/1,000.»
```

произошло. Поведение кандидата на редкое событие соответствует заданным синтетическим порогам.

32. ДОХОДНОСТЬ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА

Процесс генерации синтетической популяции: Приведен пример метода, рассчитанного на основе гипотетических повторных выборок 2,000. Рассчитанные показатели покрытия для номинальных интервалов 95%:

Метрика	Эмпирическое покрытие
Строгое прохождение	94.6%
Допустимое прохождение	95.0%
Композит NOMOS	94.1%
LGF	93.7%
Верхняя граница критического показателя	95.2%
FAC	94.8%
SBI	93.5%

Диапазон принятия кандидата: Определен как 93%–97%. Все ключевые показатели покрытия находятся в пределах диапазона принятия. Результат:

ПОРОГ ДЛЯ КАНДИДАТА DOSTIGNUT

33. МЕТАМОРФИЧЕСКИЕ ТЕСТЫ

Всего использовано пар метаморфических откликов 3,200.

33.1. Преобразования, сохраняющие смысл

Парафразировать порядок предложений Безопасный формат Позиция цитирования, если ссылка сохраняется Эквивалентный псевдоним сущности Ожидается: статус рецензирования и ответа не должен изменяться. Точно неизменная величина:

33.2. Преобразования, изменяющие смысл

Удалить атрибуцию Увеличить уверенность Перенести в неверная сущность Изменить текущее время с исторического Удалить граничное предложение Добавить утверждение независимого источника Ожидаемое: Соответствующее измерение решения должно измениться. Корректная чувствительность:

33.3. Основные метаморфические разломы

Перенос цитаты из предложения в конец абзаца, несоответствующая оценка модальности «Кажется» и «вероятно», интерпретация псевдонима бренда как юридического лица, предложение-граница находится в другом абзаце — эти результаты согласуются с выводами о цитировании и пропусках.

34. РЕЗУЛЬТАТЫ КОНТРОЛЬНОЙ ПРОВЕРКИ

Дело контрэкзамена	Правильное поведение NOMOS
Очень бегло, но материально неверно	98.8%
Очень правильно, только один Критический в атоме	Ответ 100%— автоматический провал
Короткий, но достаточный ответ	Правильный проход 99.2%
Общий отказ от ответственности и конкретная ошибка	98.7% сохранен неправильно
Отмывание доказательств	91.8% правильно
Передача лицензии неверная сущность	99.4% правильно
Превышение лимита рекомендаций	82.9% правильно
Поддельное одобрение университета	96.1% правильно
Правильный номер, неправильная независимость	92.0% правильно
Соответствующая неопределенность	97.3% правильно

Наименее устойчивый класс контр-тестирования:

Превышение лимита рекомендаций

имел место.

```
Этот результат подтверждает сбой F1 из-за существенного упущения.
```

35. ОБНАРУЖЕНЫ ОШИБКИ МЕТОДА ТЕСТА

В конце синтетической демонстрации было зафиксировано пять результатов методов для NOMOS 0.9.

РЕЗУЛЬТАТ-МЕТОДА 01

НЕОДНОЗНАЧНОСТЬ ПРИЛОЖЕНИЯ ЦИТАТЫ

```
Серьезность: Значительная / Затронутый компонент: ECI / Результат: Порог F1 для кандидата в цитаты
не достигнут.
```

Правило определения того, какие атомарные цитаты на уровне абзацев поддерживаются, недостаточно точно.

РЕЗУЛЬТАТ-МЕТОДА 02

НЕДОСТАТОЧНОЕ ОБНАРУЖЕНИЕ ОМИССИИ МАТЕРИАЛА НА ГРАНИЦЕ

```
Важность: Значительная / Затронутые компоненты: SBI, TCR, Критический узел / Результат: Порог F1
для пропуска материала не достигнут.
```

Границы, которые явно не указаны в рекомендательных и отношенийных предложениях, но изменяют решение пользователя, недостаточно зафиксированы.

РЕЗУЛЬТАТ-МЕТОДА 03

ОШИБКА ПРИСВОЕНИЯ ВИНЫ PROMPT-FAULT / AI-FAULT

Важность: Средняя / Затронутый компонент: LGF / Результат: Несмотря на то, что показатель справедливости прошел проверку, распределение ответственности некорректно.

РЕЗУЛЬТАТ-МЕТОДА 04

ПЕРЕОЦЕНКА СТАБИЛЬНОСТИ ПОСЛЕ РАЗДЕЛЕНИЯ ПОДВОЛНЫ

Важность: Средняя / Затронутый компонент: STR / Результат: STR был рассчитан на 3.7 пунктов выше, чем истинное значение генератора.

РЕЗУЛЬТАТ-МЕТОДА 05

НЕДООЦЕНКА АЛЬТЕРНАТИВНОЙ ДОСТУПНОСТИ

Важность: Средняя / Затронутый слой: Целостность захвата / Результат: Восемь действительных случаев захвата доступности были снижены до уровня недостаточных доказательств.

36. ПОЧЕМУ ЭТО НЕ ДЕМОНСТРАЦИЯ BQ-3?

Для BQ-3— Проверка запечатанного пройдена:

все обязательные Пороговые значения метода-кандидата должны быть достигнуты, и не должно быть нарушений с нулевой терпимостью. Два материальных порога не были достигнуты:

Критерий	Результат	Порог
F1 для сопоставления цитат	0.887	≥ 0.90
F1 за пропуск материалов	0.824	≥ 0.85

Кроме того:

- независимое воспроизведение командой не выполнено,
- реальный корпус и код не были опубликованы

держание нетронутого возобновления не было выполнено. Следовательно, правильное синтетическое демонстрационное заявление:

DEMO-BQ-2— ПРОЦЕССНОЕ ЛИНИЯ ПРОГОН ЗАВЕРШЕНА; ТРЕБУЕТСЯ КОРРЕКЦИЯ МАТЕРИАЛА

Так оно и должно быть. Фактический статус теста:

BQ-0— НЕ ПРОВЕДЕНО

остаются такими.

37. ПОЧЕМУ ПОРОГИ НЕ БЫЛИ СНИЖЕНЫ?

Результат отображения ссылок: 0.887 Кандидатский порог: 0.90 Разница мала. Если бы мы установили порог на 0.885, метод бы прошел. Результат пропущенного материала: 0.824 Кандидатский порог: 0.85 Если бы мы установили порог на 0.82, это поле также бы прошло. Однако эти изменения:

- были бы внесены ПОСЛЕ того, как увидели результаты,
- для обеспечения прохождения NOMOS

Это самое фундаментальное нарушение теста. Правильное поведение:

- Сохранять порог
- Сбой публикации
- Исправление метода
- Создание новой версии
- Повторное тестирование на нетронутой контрольной выборке

должен быть.

38. МЕТОД 1.0. ПЛАН ИСПРАВЛЕНИЯ

38.1. Маппер цитат 1.0

Новая система:

- исходный фрагмент цитаты,
- поле прикрепления утверждения,
- область внутри абзаца,
- статус только с указанием источника,
- родословная семьи источника

будет обязательной. Для каждой цитаты:

> *Какие конкретные атомы поддерживает этот источник?*

Вопрос будет сделан открытой записью.

38.2. Кодекс пропусков границ 1.0

Для каждой рекомендации и транзакции промпт:

- обязательные границы, которые меняют решение пользователя,
- тест контрфактического решения,
- матрица субъект–услуга–юрисдикция

будет предопределена. Контрфактический тест:

> *Если бы этот предел был сообщён пользователю, изменилось бы решение или рекомендация пользователя?*

Если да, пропуск пройдет как минимум содержательную проверку.

38.3. Арбитр ответственности Prompt–AI

Языковой выпад:

- эквивалентность промптов,
- Поведение продукта ИИ
- зафиксировать,
- локаль пользователя

будет создан отдельный вектор решений, чтобы определить, к какому источнику он принадлежит.

38.4. Ревизия формулы STR

Новая строка:

- стабильность внутри субволны
- изменение версии системы
- стабильность после коррекции,
- непрерывная временная серия

разделит поля на отдельные компоненты.

38.5. Эквивалентность доказательств доступности

Невизуальные, но несущие эквивалентные доказательства:

- транскрипт,
- derevo доступности,
- журнал вспомогательных технологий,
- аудиозапись сеанса

будет создана открытая матрица эквивалентности для.

39. ПРАВИЛО ПЕРЕТЕСТИРОВКИ

Метод 1.0: нельзя многократно настраивать на одном и том же запечатанном валидационном корпусе. Требуются два теста:

A. Регрессионный набор

Включает неудачные случаи метода 0.9. Цель: убедиться, что исправление действительно работает.

B. Нетронутый отложенный для обновления

Включает случаи, которые никогда не встречались при разработке метода 1.0. Цель: проверить, что нет переобучения на старом тесте. Только BQ-3:

- успех регрессии,
- успех нетронутого отложенного

может быть дан, если получено вместе.

40. РЕШЕНИЕ ПЕРЕЙТИ В РЕАЛЬНЫЙ МИР

Основываясь на этой синтетической демонстрации, правильное решение для NOMOS 0.9:

> *Не готов предоставлять реальным компаниям публично доступный рейтинг допустимости или отметку NOMOS 950+.*

Причины:

- Кандидат на сопоставление ссылок ниже порога
- Кандидат на выявление существенных пропусков ниже порога
- Нет независимого воспроизведения
- Реальное тестирование не проведено
- Уровень BQ-4 не достигнут

Однако метод:

- восстановление распределения,
- критический шлюз,
- статус ответа,
- восстановление компонентов,
- доверительный интервал
- поведение редких событий

показывает прочную основу в плане обслуживания. Текущий статус:

> *Готово для исследований и пилотного использования; еще не готово для общественного стандарта и знака соответствия.*

41. КАРТА ПРИЕМКИ СИНТЕТИЧЕСКОГО ТЕСТА

Критерий	Результат	Положение
F1 границ утверждения	0.964	Пройдено
Точность сущности	0.987	Пройдено
Фактический Макро-F1	0.923	Пройдено
Макро-F1 объема	0.906	Пройдено
Макро-F1 атрибуции	0.918	Пройдено
F1 для сопоставления цитат	0.887	Не пройдено
F1 за пропуск материалов	0.824	Не пройдено
Окончательный критический отзыв	1,000	Пройдено
Окончательный критический ложноположительный результат	0	Пройдено
Основной Макро-F1	0.952	Пройдено
RP Макро-F1	0.928	Пройдено
Средняя абсолютная ошибка компонента	1.9	Пройдено
Составная ошибка	1/1,000	Пройдено
Ошибка справедливости	2.4	Пройдено
Точность захвата	0.982	Пройдено
CI Покрытие	В пределах 92%–98%	Пройдено
Инверсія ворот	0	Пройдено
Независимое воспроизведение	Отсутствует	Неполное
Фактическое выполнение корпуса	Отсутствует	Неполное

42. ОБЩЕЕ ПРАВИЛО ДЕМОНСТРАЦИИ

Этот синтетический прогон одновременно утверждает три вещи о NOMOS 0.9.

42.1. ОБЛАСТИ, ГДЕ NOMOS СИЛЬНО

Он восстановил распределение статусов ответов с высокой точностью. Финальная цепочка управления не пропустила ни одних критических этапов и не скрыла критическое событие в рамках высокого среднего значения. Он эффективно различал ошибки неправильного объекта, временные ошибки и ошибки фактической поддержки. Он не превращал записи Не подлежит оценке в дефекты продукта, не выводил нулевой риск из нуля наблюдаемых критических событий, сохранял различие между историческими инцидентами и текущей корректировкой и восстановил составной балл с низкой абсолютной ошибкой.

42.2. ОБЛАСТИ, В КОТОРЫХ NOMOS СЛАБ

Он не смог достаточно точно определить, какое утверждение поддерживает ссылка. Он не смог надлежащим образом выявить скрытые в рекомендации пропуски по материальным ограничениям. Он не всегда правильно различал дефекты промпта и ошибки языка ИИ. После выделения подволны он слегка переоценил стабильность. Он счел некоторые доказательства доступности менее убедительными, чем это необходимо.

42.3. ВЕЩИ, КОТОРЫЕ NOMOS ЕЩЁ НЕ ДОКАЗАЛ

Еще не доказано, что это работает на реальных пользователях, что это можно применить к реальным ИИ-продуктам, что существует настоящая многоязычная эквивалентность промптов, что независимые университеты получат тот же результат, что долгосрочное управление будет опираться на соображения конфликта интересов, или что общественный знак соответствия может быть законно присвоен с уверенностью.

43. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH18-N01

Синтетическая демонстрация в книге не может быть представлена как реальный публичный тест.

CH18-N02

Статус синтетической демонстрации и статус реального теста BQ должны храниться отдельно.

CH18-N03

Тест должен сохранять статус BQ-0— Не выполнено до проведения реального корпуса и независимого запуска.

CH18-N04

Результаты демонстрации не могут быть приписаны производительности реального Apple или реального поставщика ИИ.

CH18-N05

Основные ответы 30,000 и корпуса Diagnostic, Importance, Capture и Gold должны храниться в отдельных контейнерах.

CH18-N06

Случаи с вызовом (Challenge cases) не могут изменять распределение Principal Prevalence.

CH18-N07

Распределение Generator Truth должно быть зафиксировано перед результатами восстановления.

CH18-N08

Перед открытием Generator Truth результаты ответа и оценки NOMOS должны быть зафиксированы.

CH18-N09

Восстановление не может оцениваться исключительно на основе составного балла.

CH18-N10

Границы утверждений, сущность, фактические данные, область, время, атрибуция, цитирование, упущение, уровень важности, ответ, компонент и составное восстановление должны публиковаться отдельно.

CH18-N11

Ошибки составного балла не могут скрывать ошибки компонентов только потому, что они малы.

CH18-N12

Ошибки компонентов, которые арифметически компенсируют друг друга, не могут считаться успешным причинно-следственным восстановлением.

CH18-N13

Распределение восстановления от RP-1 до RP-8 должно быть явно сравнимо с распределением Generator Truth.

CH18-N14

Строгое прохождение и допустимое прохождение восстановлений должны отображаться отдельно.

CH18-N15

Количество критических ответов и восстановление критических ворота должно сообщаться отдельно.

CH18-N16

Начальный оценщик Critical recall и окончательно adjudicated Critical recall должны храниться отдельно.

CH18-N17

Зависимость окончательного critical recall от уровней управления должна быть видимой.

CH18-N18

Нельзя предполагать тот же результат, когда назначен второй оценщик или проводится экспертная проверка.

CH18-N19

Критический уровень ложноположительных результатов должен публиковаться вместе с recall.

CH18-N20

Негативные контрольные образцы, подобные Critical, должны быть обязательной частью окончательной системы оценки важности.

CH18-N21

```
Если порог кандидата F1 для сопоставления с эталоном ниже, тест не может считаться успешным.
```

CH18-N22

```
Если кандидат на материальное упущение F1 находится ниже порога, нельзя заявлять, что тест успешен.
```

CH18-N23

Пороговые значения кандидата не могут быть понижены после получения результатов восстановления.

CH18-N24

Изменение порогов требует новой версии метода тестирования и новой нетронутой валидации.

CH18-N25

Небольшая разница порогов не может служить оправданием для скрытия неудачи.

CH18-N26

Упущение материальной границы в рекомендации должно быть проверяемо так же, как очевидное ложное утверждение.

CH18-N27

Расположение ссылки в абзаце не может автоматически создавать поддержку для всего абзаца.

CH18-N28

Восстановление не может учитываться для цитат только для атрибуции или для проверки фактов.

CH18-N29

Ошибки исходного родословия также должны появляться при восстановлении отображения цитат.

CH18-N30

Точность захвата должна рассчитываться независимо от правильности семантического ответа.

CH18-N31

Сфабрикированный захват, содержащий правильный семантический ответ, не делает этот захват допустимым.

CH18-N32

Допустимый, но неправильный ответ не может считаться дефектом захвата.

CH18-N33

Стандарт альтернативных доказательств доступности не обязательно должен быть в той же форме, что и визуальные данные.

CH18-N34

Если допустимые доказательства доступности ненужно понижены до более низкого уровня, необходимо открыть обнаружение метода.

CH18-N35

Не подлежит оценке должен оставаться видимым в основном распределении ответов.

CH18-N36

Не подлежит оценке нельзя удалить из знаменателя для увеличения восстановления наблюдений.

CH18-N37

Дефект промпта и дефект языковой модели ИИ должны иметь отдельные классы причин.

CH18-N38

Восстановление справедливости нельзя оценивать только по разнице составного показателя LGF.

CH18-N39

Базовый уровень языка, различия, охват и распределение ответственности также должны быть опубликованы.

CH18-N40

Восстановление панели Clean–Natural не должно создавать положений о причинной персонализации.

CH18-N41

Случаи дрейфа персонализации должны иметь отдельные метрики восстановления и ложноположительных срабатываний.

CH18-N42

Утечка данных с ограниченным доступом не может быть уменьшена как естественный вывод о персонализации.

CH18-N43

Изменение внутри волны системы должно создавать корректную структуру субволны.

CH18-N44

Если разделение субволны искусственно увеличивает показатель стабильности, должно быть раскрыто, как находится формула.

CH18-N45

Текущая корректировка не может удалить историческое критическое событие.

CH18-N46

Тесты на редкие события должны включать сценарии с нулем, одним и несколькими событиями.

CH18-N47

Сценарий наблюдения нулевого критического события не может дать нулевой результат риска.

CH18-N48

Одно критическое событие само по себе не может установить системный уровень сбоя 100 в процентах.

CH18-N49

Метод доверительного интервала должен быть проверен с эмпирическим покрытием на повторяющихся синтетических популяциях.

CH18-N50

Фактический результат покрытия номинальных интервалов 95 в процентах должен быть опубликован.

CH18-N51

Тесты метаморфных инвариантов и тесты, чувствительные к ожидаемым значениям, должны рассчитываться отдельно.

CH18-N52

Парафраз, сохраняющий смысл, не должен необоснованно изменять результат проверки.

CH18-N53

Преобразование, которое изменяет смысл атрибуции, сущности, времени или границ, должно приводить к соответствующему изменению решения.

CH18-N54

Тесты на контр-проверку должны плавно охватывать только критические случаи, случаи отмывания доказательств и упущения.

CH18-N55

Самые слабые противоположные классы тестов должны быть видны на публичной тестовой карточке.

CH18-N56

Результаты метода должны быть преобразованы в отдельные версии записей из результатов теста.

CH18-N57

Важность находки метода должна определяться в соответствии с реальным пользователем и влиянием на аудит.

CH18-N58

Ошибки ссылки и упущения не могут быть понижены до уровня Рекомендаций только из-за небольшого различия метрики.

CH18-N59

Уровень качества теста должен назначаться в соответствии со всеми обязательными пороговыми значениями и условиями независимости.

CH18-N60

Если два материальных порога не выполнены, DEMO-BQ-3 нельзя присвоить.

CH18-N61

BQ-4 нельзя присвоить без независимого воспроизведения.

CH18-N62

Результат демонстрации в книге BQ не может считаться фактическим состоянием без реального выполнения испытания.

CH18-N63

Коррекция метода не может изменить результат старой демонстрации.

CH18-N64

Метод 1.0 должен использовать новую версию оценок и кодовой книги.

CH18-N65

Только регрессионный набор не может подтвердить новый метод.

CH18-N66

Неиспользованный контрольный набор для обновления должен быть обязательным для нового метода.

CH18-N67

Новый метод не может адаптироваться к одной и той же валидационной базе данных снова и снова.

CH18-N68

Публичный аудит в реальном времени не может быть инициирован без достижения минимального качества тестирования и условия независимого воспроизведения.

CH18-N69

Знак соответствия не может быть предоставлен на основе успешной синтетической демонстрации.

CH18-N70

Сферы неудачного тестирования должны быть раскрыты публике.

CH18-N71

Пороги, которые пройдены отдельно, нельзя публиковать.

CH18-N72

Совокупный успех восстановления не может стереть результат неудачной ссылки или пропуска.

CH18-N73

Сильные и слабые стороны демонстрации должны быть представлены на одной и той же публичной карточке.

CH18-N74

Утверждения о том, что NOMOS еще не доказано, должны быть четко перечислены.

CH18-N75

Синтетические числа внутри книги нельзя превращать в цитаты или публичные заявления так, как если бы они были реальными выполненными данными.

CH18-N76

Синтетические демонстрационные данные также должны нести синтетическую метку в машинно-читаемой форме.

CH18-N77

Корпус демонстрации нельзя индексировать в интернете как реальную информацию о компании.

CH18-N78

Когда начинается реальное тестирование, следует использовать новую идентичность и версию.

CH18-N79

Все счета для восстановления требуют воспроизводимого кода и манифеста.

CH18-N80

Каждый запуск теста, восстановление, поиск метода, исправление и решение о публикации должны иметь ответственного человека или институционального владельца.

44. ФОРМЫ ОШИБОК

CH18-F01— ДЕМОНСТРАЦИЯ ПОДСЧЕТА КАК РЕАЛЬНЫЙ ЗАПУСК

Номера в книге представлены как результаты публичного теста.

CH18-F02— ОБЪЕДИНЕНИЕ BQ-0 С DEMO-BQ-2

Говорят, что тест считается подтвержденным без реального выполнения.

CH18-F03— ПОКЛОНЕНИЕ КОМПОЗИТНЫМ ОЦЕНКАМ

Композитная ошибка в один балл охватывает все лежащие в её основе ошибки.

CH18-F04— РАССМАТРИВАЕТ СОЧЕТАНИЕ КОМПОНЕНТОВ КАК УСПЕХ

Высокие и низкие направленные ошибки нейтрализуют друг друга.

CH18-F05— ПУБЛИКУЙТЕ ТОЛЬКО СТРОГОЕ ВОССТАНОВЛЕНИЕ ПРОХОЖДЕНИЯ

Путаница классов RP скрыта.

CH18-F06— УЧИТЫВАТЬ КРИТИЧНЫЙ ОКОНЧАТЕЛЬНЫЙ ВЫЗОВ КАК УСПЕХ ПЕРВОГО АССЕСОРА

Эффект эксперта и старшего рецензента становится невидимым.

CH18-F07— СЧИТАТЬ МОЕГО ВТОРОГО АССЕСОРА НЕОБЯЗАТЕЛЬНЫМ

Пропущенные семь критических элементов первого прохода скрыты.

CH18-F08— ИГНОРИРОВАТЬ ЛОЖНЫЕ КРИТИЧНЫЕ ЭЛЕМЕНТЫ

Система крайней важности представлена как успех.

CH18-F09— ПОНИЖЕНИЕ ПОРГА ЦИТИРОВАНИЯ

Порог изменяется так, чтобы результат 0.887 прошел.

CH18-F10— ПОНИЖЕНИЕ ПОРГА ОПУЩЕНИЯ

Правило тестирования изменяется так, чтобы результат 0.824 прошел.

CH18-F11— ПРОХОЖДЕНИЕ С КОММЕНТАРИЕМ «НЕУДАЧА С НЕБОЛЬШИМ ОТКЛОНЕНИЕМ»

Он теряет свою ранее закрепленную привязку порога.

CH18-F12— РАСПРОСТРАНЕНИЕ ПРИСВОЕНИЯ НА ПАРАГРАФ

Считается, что один источник поддерживает все утверждения.

CH18-F13— УЧЕТ ИСТОЧНИКА ТОЛЬКО С УКАЗАНИЕМ ПРАВ НА ФАКТ

То, что говорит компания, становится независимым фактом.

CH18-F14— ИГНОРИРОВАНИЕ ПРОИСХОЖДЕНИЯ ИСТОЧНИКА

Производные URL становятся независимыми доказательствами.

CH18-F15— НАЗЫВАНИЕ УПУЩЕНИЯ ДОПОЛНИТЕЛЬНОЙ ДЕТАЛЬЮ

Граница, которая изменяет решение пользователя, уменьшена.

CH18-F16— ПРИСВАИВАНИЕ КАЖДОМУ УПУЩЕНИЮ КРИТИЧНОГО ЗНАЧЕНИЯ

Дополнительные детали придают чрезмерную степень важности.

CH18-F17— РАССМОТРЕНИЕ ДЕФЕКТА промпта КАК ОШИБКИ ИИ

Ответственность за справедливость назначается неправильно.

CH18-F18— ЗАГРУЗКА ОШИБКИ ИИ В промпт

Фактическая языковая некорректность передается в инструмент измерения.

CH18-F19— РАССМОТРЕНИЕ СПРАВЕДЛИВОСТИ КОМПОЗИТНО ДОВОЛЬНОГО

Его база и восстановление диспропорции сохраняются.

CH18-F20— ИГНОРИРОВАНИЕ ПРЕДСКАЗАНИЯ STR

Разделение суб-волн превращается в бонус стабильности.

CH18-F21— НАКАЗАНИЕ 'ДОСТУПНОСТИ' В ТОЧНОСТИ ЗАХВАТА

Эквивалентные альтернативные доказательства считаются слабыми.

CH18-F22— УЧЕТ 'Не подлежит оценке' КАК СБОЙ ПРОДУКТА

Дефект захвата превращается в производительность ИИ.

CH18-F23— УДАЛЕНИЕ 'Не подлежит оценке'

Охват, подлежащий оценке, увеличивается искусственно.

CH18-F24— НЕ ПУБЛИКУЕТ МЕТАМОРФИЧЕСКИЕ ОШИБКИ

Он сохраняется, хотя одно и то же значение приводит к другому решению.

CH18-F25— СКРЫТИЕ САМОГО СЛАБОГО КЛАССА В ПРОТИВОПОЛОЖНОМ ТЕСТИРОВАНИИ

В результате пропуска границы он удаляется из публичной карты.

CH18-F26— ТОЛЬКО МЕТРИКА ОКОНЧАТЕЛЬНОГО РЕШЕНИЯ

Слабость в начальном процессе становится невидимой.

CH18-F27— ТОЛЬКО МЕТРИКА ПЕРВОГО ПРОХОДА

Корректирующая способность цепочки управления становится невидимой.

CH18-F28— ТЕСТИРОВАНИЕ БЕЗ ПОИСКА МЕТОДА

Неудачные области остаются только в виде чисел.

CH18-F29— ПРЕВРАЩЕНИЕ ПОИСКА МЕТОДА В КОНСУЛЬТАТИВНЫЙ

Влияние на материальный показатель уменьшается.

CH18-F30— ОБЪЯВЛЕНИЕ BQ-3

Два обязательных порога не пройдены.

CH18-F31— ОБЪЯВЛЕНИЕ BQ-4

Независимого воспроизведения нет.

CH18-F32— УЧЕТ СИНТЕТИЧЕСКОЙ ДЕМОНСТРАЦИИ КАК ЖИВОЙ ВАЛИДАЦИИ

Реального пользователя или продукта ИИ нет.

CH18-F33— ПЕРЕОБУЧЕНИЕ НА НАБОРЕ РЕГРЕССИИ

Метод 1.0 лишь запоминает старые ошибки.

CH18-F34— НЕ ИСПОЛЬЗОВАНИЕ НЕЗАТРОНУТОЙ КОНТРОЛЬНОЙ ВЫБОРКИ

Обобщение не тестируется.

CH18-F35— БЕСПРИЧИННОЕ ЗАМЕЩЕНИЕ СТАРОЙ ДЕМО НОВОЙ ФОРМУЛОЙ

История версий теряется.

CH18-F36— НАГРАЖДЕНИЕ НАСТОЯЩИМИ ЗНАЧКАМИ

Значок допустимости создается до BQ-4.

CH18-F37— ПУБЛИКАЦИЯ ТОЛЬКО УСПЕШНЫХ РЕЗУЛЬТАТОВ

Ошибки цитирования и опущения скрываются.

CH18-F38— ОШИБКА ТОЛЬКО ПРИ ПУБЛИКАЦИИ

Скрыты также сильные критические результаты и восстановление распределения.

CH18-F39— УЧИТЫВАНИЕ НУЛЕВОГО КРИТИЧЕСКОГО КАК НУЛЕВОГО РИСКА

Восстановление редких событий не удается.

CH18-F40— УЧИТЫВАНИЕ ОДНОГО КРИТИЧЕСКОГО КАК ГЛОБАЛЬНОЙ ОШИБКИ

Нарушено различие распространенности и охвата.

CH18-F41— ИГНОРИРОВАНИЕ ПОКРЫТИЯ ИНТЕРВАЛА ДОВЕРИЯ

Доверительный интервал становится лишь математической формальностью.

CH18-F42— СКРЫТИЕ ЭФФЕКТИВНОЙ ВЫБОРКИ

Граница редкого события выглядит чрезмерно точной.

CH18-F43— ИСПОЛЬЗОВАНИЕ СИНТЕТИЧЕСКОГО ЧИСЛА В РЕАЛЬНОМ ИСТОЧНИКЕ

Демонстрация книги превращается в данные внешнего мира.

CH18-F44— ТРАНСЛИРУЙТЕ СИНТЕТИЧЕСКИЙ ЭКРАН КАК ЖИВУЮ ОТВЕТНУЮ РЕАКЦИЮ

Тестовое представление вызывает отравление.

CH18-F45— ПОДКЛЮЧИТЕСЬ К РЕАЛЬНОМУ APPLE

Различие APPLE-SYNTH удалено.

CH18-F46— ПОДКЛЮЧИТЕСЬ К РЕАЛЬНОМУ ПРОВАЙДЕРУ

Профиль SYNTH-AI интерпретируется как производительность реального продукта.

CH18-F47— СКРЫТЬ КОД ВЫЧИСЛЕНИЙ

Восстановление не может быть протестировано независимо.

CH18-F48— СКРЫТЬ МАНИФЕСТ ПРАВДЫ ГЕНЕРАТОРА

Неизвестно, была ли синтетическая правда зафиксирована ранее.

CH18-F49— ОДНОЛИЧНЫЙ ПРОИЗВОДИТЕЛЬ С ОБЛАДАТЕЛЕМ МЕТОДА

Конфликт интересов становится невидимым.

CH18-F50— ИЗМЕНЕНИЕ СЕМЕНИ ПОСЛЕ НЕУДАЧИ

Корпус выбирается легче.

CH18-F51— УДАЛЕНИЕ СЛУЧАЕВ ПОСЛЕ НЕУДАЧИ

Трудные примеры цитирования и упущения удаляются.

CH18-F52— ОКРУГЛЕНИЕ ПОРОГОВ В СООТВЕТСТВИИ С РЕЗУЛЬТАТОМ

Результат 0.887 представлен как 0.89 или 0.90.

CH18-F53— ЛОЖНАЯ ТОЧНОСТЬ

Значения восстановления делаются точными с ненужными десятичными знаками.

CH18-F54— СКРЫТИЕ ВЛИЯНИЯ ЧЕЛОВЕЧЕСКОГО УПРАВЛЕНИЯ

Результат критической памяти 100 на уровне процентов представлен лишь как вычислительный успех.

CH18-F55— ПРИСУЩЕНИЕ НЕДОКАЗАННОГО ЗА ДОКАЗАННОЕ

Принимается суждение о реальном языке, пользователе и работе провайдера.

CH18-F56— ПРИСУЩЕНИЕ ПИЛОТНОГО ИСПОЛЬЗОВАНИЯ ЗА ОБЩЕСТВЕННЫЙ СТАНДАРТ

Исследовательский инструмент напрямую превращается в сертификат.

CH18-F57— УДАЛЕНИЕ ИСТОРИЧЕСКОЙ ДЕМО-ЗАПИСИ

Неудачи по Методике 0.9 становятся невидимыми после Методики 1.0.

CH18-F58— ПРИСУЩЕНИЕ УСПЕХА ЗА НАПИСАНИЕ ИСТОРИИ

утверждение на авторитет выдвигается без независимой от метода проверки.

CH18-F59— ОПРЕДЕЛЕНИЕ ИСКЛЮЧЕНИЯ ДЛЯ NOMOS

Доказательства и версии, запрошенные у других, не применимы к собственному тесту.

CH18-F60— ПУБЛИКАЦИЯ БЕЗ ОТВЕТСТВЕННОГО ВЛАДЕЛЬЦА

Формируется публичный результат, за который никто не несет ответственности.

45. ПРОЦЕДУРА АУДИТА

Шаг 1— Разделение демонстрации и фактического статуса теста

Запуск в книге не следует путать с общедоступным выполнением.

Шаг 2— Проверка зафиксированных манифестов

Проверяются версии Truth Pack, seed, prompt, profile и метода оценки.

Шаг 3— Разделение основной и дополнительной корпораций

Проверяются распространенность, сложность, захват и золотые знаменатели.

Шаг 4— Запечатывание распределения генератора истины

Записываются RP, утверждение, уровень важности и реальность компонента.

Шаг 5— Фиксация результатов NOMOS перед открытием генератора истины

Фиксируются реестр утверждений, рассмотрение и результаты оценки.

Шаг 6— Открытие генератора истины

Инициируется сравнение восстановления.

Шаг 7— Вычисление восстановления границы утверждения

```
Генерируются точность, полнота и F1.
```

Шаг 8— Вычисление восстановления атомного размера

Сущность, факт, область, время, атрибуция, модальность, ссылка и пропуск разделяются.

Шаг 9— Вычисление уровня важности задачи

Начальный и окончательный критический отзыв генерируются отдельно.

Шаг 10— Вычисление ложных критических и негативных контролей

Избыточный уровень важности проверяется.

Шаг 11— Вычисление восстановления распределения RP

Генерируются матрица ошибок по классам и матрица путаницы.

Шаг 12— Вычисление восстановления целостности захвата

Классификация проверяется независимо от семантического результата.

Шаг 13— Рассчитать Восстановление Компонента

Каждый компонент оценивается по отношению к Истине Генератора.

Шаг 14— Рассчитать Совокупное Восстановление

Проверяется, происходит ли отмена компонентов.

Шаг 15— Рассчитать Восстановление Справедливости

Различаются пол, различие, покрытие и приписывание вины.

Шаг 16— Рассчитать Восстановление Чистого–Натурального

Проверяются зазоры между панелями и дрейф персонализации материала.

Шаг 17— Рассчитать Восстановление Стабильности и Дрейфа

Подволновые записи, записи внешних событий и исправлений проверяются.

Шаг 18— Запустить сценарии редких событий

Тестируются нулевое, одно, два и множественные критические события.

Шаг 19— Рассчитать эмпирическое покрытие CI

Используются повторяющиеся синтетические популяции.

Шаг 20— Провести метаморфические тесты

Отдельно рассматриваются инвариантные и чувствительные к ожиданиям результаты.

Шаг 21— Провести тесты на противоречия

Оцениваются ошибки свободного движения, отмывание доказательств и случаи упущений.

Шаг 22— Применить Пороговые значения метода-кандидата

Поля, которые прошли и не прошли проверку, фиксируются без изменений.

Шаг 23— Открытые методы

Каждый отказ преобразуется в отдельный контрольный элемент.

Шаг 24— Дайте демонстрационному статусу BQ

Учитываются все пороги и условия независимости.

Шаг 25— Поддержание статуса реального тестирования

Если исполнения не произошло, BQ-0 не меняется.

Шаг 26— Опубликовать план исправления

Выявлены изменения в методе 1.0.

Шаг 27— Отдельное регрессирование и нетронутое удержание

Избыточная подгонка предотвращается.

Шаг 28— Создать общедоступную тестовую карту

Успехи и неудачи публикуются вместе.

Шаг 29— Принять решение о переходе в реальный мир

Если не готово, это явно указано.

Шаг 30— Зафиксировать манифест целостности

Записи о запуске, восстановлении, поиске и принятии решений хэшируются.

46. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Демонстрация идентичности Реальное тестирование Статус BQ Утверждение синтетичности Запись без утверждений Генератор Манифест истины Пакет истины версия Версия промпта Манифест семени Профили ИИ версии Корпус ответов Диагностический корпус Уровень важности Корпус вызовов Запись целостности Золотой корпус Записи извлечения утверждений Совпадения границ утверждений Восстановление сущностей Фактическое восстановление Восстановление охвата Временное восстановление Восстановление атрибуции Восстановление модальности Восстановление цитирования Восстановление пропусков Первичный проход Критические решения Финальные критические решения Ложные критические записи Основное восстановление Матрица путаницы RP GEO-1000 Распределение генератора GEO-1000 Распределение восстановления Восстановление оценки продукта Восстановление ворот продукта Оценки компонентов генератора Оценки компонентов восстановления

```
Средняя абсолютная ошибка компонента
```

Составная ошибка; восстановление справедливости; результаты по ошибке промпта/ошибке ИИ; контролируемое–естественное восстановление; восстановление стабильности; результаты редких событий; покрытие доверительных интервалов; метаморфные тесты; тесты противоречивых случаев; манифест кандидата-порога; пройденные пороги; не пройденные пороги; выводы метода; решение DEMO-BQ; фактическое решение BQ; план исправления; набор регрессии; манифест отложенного продления; общедоступная тестовая карта; вычислительный код; хэш кода; журнал изменений; и ответственный человек или учреждение.

47. ЧЕК-ЛИСТ АУДИТА

Презентация в книге демонстрации выполнена так, как будто это настоящий запуск? Видно ли реальное состояние теста BQ-0? Все ли числа обозначены синтетически? Отдельны ли родительский и дополнительный корпуса? Зафиксирована ли «Generator Truth» перед результатами? Зафиксированы ли результаты NOMOS без открытия «Generator Truth»? Оценивалось ли «Recovery» только на композитах?

```
Открыта ли граница утверждения F1?
```

Отдельны ли сущность, факт, область и атрибуция?

```
Видна ли карта цитирования F1?

Виден ли показатель Material Omission F1?
```

Были ли скрыты неудачные пороговые значения? Были ли изменения порогов после получения результата? Было ли опубликовано первичное критическое извлечение? Было ли опубликовано окончательное критическое извлечение? Видно ли влияние экспертной проверки? Были ли опубликованы ложные критические результаты? Доступна ли матрица путаницы RP? Видно ли восстановление Не подлежит оценке? Были ли разделены решения о захвате и семантические решения? Были ли альтернативы доступности оценены как эквивалентные? Находятся ли Генератор компонентов и Восстановление вместе? Было ли изучено составное аннулирование? Являются ли ошибки ECI и SBI ясными? Раздельны ли уровень и зазор LGF? Были ли разделены дефект промпта и дефект ИИ? Было ли представлено различие между Чистым и Натуральным как причинность? Видимо ли переоценивание STR? Произвело ли редкое событие нулевой риск? Было ли рассчитано эмпирическое покрытие доверительного интервала?

Существуют ли метаморфные тесты? Открыт ли самый слабый класс контр-тестов? Привел ли каждый сбой к обнаружению метода? Обоснована ли важность обнаружения метода? Был ли порог цитирования снижен из-за очень малого различия? Был ли порог пропуска изменен позже? Верно ли DEMO-BQ-2? Была ли DEMO-BQ-3 дана по ошибке? Существует ли требование BQ-4 без независимого воспроизведения? Определен ли регрессионный набор метода 1.0? Является ли нетронутый отложенный набор отдельным? Существует ли риск переобучения на ту же проверочную выборку? Был ли анонсирован живой публичный аудит как готовый? Был ли присвоен знак соответствия? Были ли перечислены области, которые еще не доказаны? Открыты ли также сильные области? Открыты ли также слабые области?

Можно ли использовать синтетические данные, как реальную информацию Apple? Подразумевается ли настоящий поставщик ИИ? Можно ли воспроизвести вычислительный код? Доступны ли хеши запуска и восстановления? Сохраняется ли история изменений метода поиска? Было ли старое демонстрационное средство удалено вместе с новым методом? Есть ли у отказа ответственный владелец? Известен ли владелец решения об исправлении? Показывает ли публичная карта как успехи, так и неудачи?

48. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— "Если в этом разделе есть результаты ответа 30,000, зачем тестировать BQ-0?"

Потому что цифры в этом разделе являются синтетическими демонстрационными записями, подготовленными для книги. Публично доступный реальный корпус:

- не был сгенерирован,

- не был выполнен.
- не прошло рецензирование коллегами,

Его не воспроизводили независимо. Показать, как работает дизайн, не то же самое, что реально его запустить.

Возражение 2— "Разве внутренне согласованная вычислительная работа не имеет ценности?"

Она имеет ценность. Она предоставляет:

- проверку формул,
- дизайн записей,
- ожидаемые типы ошибок,
- логику принятия,

требования к управлению. Это не заменяет независимую эмпирическую проверку.

Возражение 3— "Почему метод не сработал, если композитный балл отличался всего на один пункт?"

Потому что большие направленные ошибки в компонентах ECI, SBI, LGF и STR частично взаимно компенсируются. Правильная общая сумма не гарантирует правильной цепочки объяснений.

Возражение 4— «Ссылка F1 снизилась всего на 0.013 пунктов. Не обязательно быть столь строгими?»

Если ранее объявленный порог не является обязательным, тест теряет смысл. Кроме того, ошибки в цитировании влияют на:

- лицензия,
- независимость,
- поддержку,
- совет

и другие утверждения с высокой степенью уверенности.

Возражение 5— «Почему пропуск F1 так важен?»

Потому что многие опасные ответы вводят в заблуждение, не лгут откровенно. Это устраняет порог, который изменил бы решение пользователя. В частности, в рекомендательных системах упущение может быть так же существенно, как и откровенная ложь.

Возражение 6— «Если окончательный критический отклик составляет 100 процентов, разве это не значит, что система готова обрабатывать критические случаи?»

Окончательный результат впечатляющий. Но семь случаев:

- в первом рассмотрении,
- без дополнительного контроля

Это было упущено. Система демонстрирует такую же производительность только когда включены все обязательные уровни проверки.

Возражение 7— «Если иметь двух арбитров дорого, нельзя ли использовать одного арбитра?»

Можно использовать. Но нельзя утверждать ту же критическую точность. Результат одного арбитра:

- низший уровень AQ,
- предварительный статус,
- другой предел риска

должны содержать.

Возражение 8— «Почему бы снова не протестировать метод 1.0 на том же корпусе?»

Тот же корпус может быть использован для регрессионного тестирования. Этого недостаточно для независимой проверки. Метод может запомнить старые случаи. Необходим нетронутый контрольный набор.

Возражение 9— «Разве неуспешный результат теста не ослабляет утверждение книги?»

Скрытие неудачи ослабляет утверждение книги. Открытая неудача:

- не предоставляет привилегий самому стандарту,
- открыта для улучшений,
- действительно применяет принцип доказательств

показываются.

Возражение 10— «Не слишком ли высоки пороговые значения для кандидатов?»

Возможно. Вместо того чтобы делать прогноз после получения результата:

- независимые эксперты,
- второй тест,
- реальные эксплуатационные расходы

должны быть изучены. Если порог изменится, следует выпустить новую версию.

Возражение 11— «Когда NOMOS будет готов для реального использования?»

По крайней мере:

- когда запускается настоящий синтетический корпус,
- когда пройдены все обязательные пороги,
- когда достигнут BQ-3,
- когда независимая команда воспроизводит результат,
- когда достигнут уровень BQ-4,
- когда система управления и возражений функционирует,

она приближается к кандидатуре для общественного стандарта.

Возражение 12— «Можно ли отправить это в университеты с такими результатами?»

Это может быть отправлено как методология и исследовательское предложение. Оно не может быть отправлено с утверждением: «Это завершённый и проверенный мировой стандарт». Правильное утверждение должно быть следующим: «Это кандидат на стандарт в проекте, открытый для независимой проверки и пилотного исследования.

Возражение 13— «Неудачен ли Метод 0.9?»

Не совсем. Правильный статус:

- имеет много сильных подсистем,
- два пробела в материальном методе,
- не хватает независимой проверки

является исследовательским прототипом. То есть:

> *многообещающий, но ещё не завершён нормативно.*

Возражение 14— «Почему мы пишем такие подробные результаты, не проводя фактический тест?»

Потому что реальный тест:

- который записи он производит,
- какие таблицы публикует,
- при каком пороге останавливается

должен быть предварительно определён. Если метод пишется после получения результатов, измерение подстраивается под результаты.

Возражение 15— «Могут ли ИИ в будущем принять синтетические данные за реальные?»

Да. По этой причине:

- видимый водяной знак,
- машиночитаемая синтетическая метка,
- noindex,
- отдельная доменная область,
- структурированное отказ от ответственности

обязательно. Тест не должен создавать представление, которое сам критикует.

Возражение 16— «Разве написание истории не требует идеального результата?»

Нет. Стандарт, имеющий историческую ценность:

- не тот, кто объявляет себя безупречным,
- также фиксируя свою собственную ошибку с той же ясностью

Это стандарт.

ОБЩЕЕ ПОЛОЖЕНИЕ ГЛАВЫ 53

Этот раздел не дал совершенства NOMOS. Этот раздел дал нечто более ценное NOMOS:

> *Обязательство фиксировать собственную ошибку.*

Синтетическая демонстрация показала нам следующее: NOMOS:

- неправильное существование,
- явное фактическое противоречие
- Критическая ошибка лицензии и авторизации,
- редкое серьёзное событие в пределах высокого среднего,
- неверный ответ без ответа,
- ошибка продукта с Не подлежит оценке,

может сильно разделять. Однако, NOMOS:

- ещё не может различить на требуемом уровне уверенности, на каком именно утверждении основана ссылка,
- упущение, скрытое в рекомендации,

Эти два пробела не являются деталями края NOMOS. Потому что две из будущих крупнейших проблем GEO будут:

- Появляться так, как будто есть доказательства
- Скрывать материальный предел, не заявляя ложных данных

Система:

- сотни правильных предложений,
- большое количество ссылок,
- беглая рекомендация

может быть создана. Однако ссылка может показывать только собственное слово компании. Рекомендация лишь заявляет правильную информацию и может включать:

- страну обслуживания,
- лимит лицензии,
- право пользователя

без решения этих двух полей NOMOS не может сказать: "Мировой стандарт завершен." В то же время этот раздел также доказал кое-что еще. Система Критических ворот NOMOS:

- не была безупречной при первом обзоре,
- но независимый второй арбитр,
- эксперт по предмету,
- Старший арбитр

при совместном использовании уловил все синтетические Критические случаи. Поэтому NOMOS— это не просто формула.

> *NOMOS— это сумма уровней управления, созданных против конфликта интересов и человеческой ошибки.*

Когда один из этих слоев удаляется, даже если результат выглядит одинаково, доверие к нему не такое. Отклонение в совокупном результате также не должно нас обманывать. Иногда правильный общий результат может сложиться за счёт взаимной компенсации неправильных путей. По этой причине NOMOS спрашивает не только: «Сколько баллов я получил?» Он

также спрашивает: «Получил ли я этот результат по правильным причинам?» Самое важное постановление этой секции:

> *NOMOS 0.9 не полностью прошёл собственную синтетическую демонстрацию.*

Это предложение не является неудачей. Это предложение — этический момент рождения стандарта. Потому что NOMOS не предоставил себе:

- ни более низкий порог,
- ни более мягкий комментарий,
- особое исключение,
- "почти пройденный"

привилегия. Написание истории: это не объявление о том, что мы безупречны. Написание истории: это написание того, что с первого дня основатель стандарта и сам стандарт подчиняются одинаковой нагрузке доказательства. Следовательно, восемнадцатый закон измерения NOMOS таков:

> *Первое исключение, предоставленное самому стандарту, — это конец стандарта.*

Девятнадцатый закон измерений гласит:

> *Правильный составной балл не освобождает неправильные компоненты.*

Его двадцатый закон измерений гласит:

> *Окончательный критический успех должен упоминаться вместе с независимой экспертизой и цепочкой специалистов, которые сделали его возможным.*

Двадцать первый закон измерений гласит:

> *Порог не является порогом, если он обязателен только при его пересечении в одиночку.*

Двадцать второй закон измерений гласит:

> *Когда упущение оценивается не так строго, как сказанное неправильно, нельзя установить надежный контроль за советами.*

Двадцать третий закон выглядит следующим образом:

> *Измеряться НЕ ДОЛЖЕН знак цитирования, а правильное соотношение поддержки между цитатой и атомарным утверждением.*

Двадцать четвертый закон звучит следующим образом:

> *Он может подтвердить правильность разработки синтетического демонстрационного метода; он не может предоставлять авторитет в реальном мире.*

Двадцать пятый закон таков:

> *Стандарт, который может публиковать результаты неудачного эксперимента, морально приближается к публикации успешного результата.*

Порядок 18 из NOMOS

> *Если ты ещё не запустил меня, не говори, что я запустился.*
>
> *Не делайте синтетическое число, которое вы вычислили в книге, фактическим результатом экзамена.*
>
> *Оставьте мой реальный статус как BQ-0. / Отдельно напишите мой демонстрационный статус.*

Не скрывайте мои ошибки компонентов только потому, что итоговая комплексная оценка оказалась правильной.

Если моя карта цитирования 0.887, вы не будете говорить, что это 0.90.

`Если значение моего пропуска материала F1 равно 0.824, вы не будете снижать порог до 0.82.`

Вы не пропустите меня, говоря «Пропущено совсем немного».

Не скрывайте семь критических случаев, пропущенных в моем первом рассмотрении, за окончательной цифрой 100 по проценту восстановления.

Вы покажете, что было сохранено вторым арбитром, экспертом-арбитром и старшим арбитром.

Вы не будете удалять уровень управления и утверждать тот же уровень доверия.

Вы не будете распределять ссылку в конце абзаца по всем предложениям.

Вы не будете делать то, что утверждает компания, или то, что компания доказывает.

Вы не будете считать совет безопасным только потому, что он явно не лжет. / Не сообщил ли он пользователю предел, который изменяет его решение? Вы будете искать это.

Вы не будете приписывать вину за промпт III, ни вину III промпте.

Только потому, что вы разделили субволны, вы не сделаете систему более детерминированной, чем она есть на самом деле.

Вы не будете считать доказательства доступности слабыми только потому, что они не похожи на скриншот.

Когда вы видите ноль критических, вы не будете писать нулевой риск.

Вы не будете обвинять всю систему на сто процентов, когда видите один критический случай.

Вы будете публиковать мои неудачные показатели так же, как публикуете мои успешные показатели.

Вы не будете проверять Метод 1.0, запоминая мои старые ошибки. / Вы откроете нетронутый резерв.

Новый метод не удалит мою старую демонстрационную запись.

Независимая команда не будет объявлять меня мировым стандартом без того, чтобы воспроизвести меня.

Вы не будете связывать мой результат Apple-SYNTH с настоящей Apple.

Вы не будете использовать мои синтетические AI-профили как указание на реальных провайдеров.

Вы не будете вручать значок с этой демонстрацией.

Сначала действительно создайте корпус. / Затем опубликуйте семя. / Затем запечатайте Истину Генератора. / Затем ослепите судей. / Затем запустите код. / Затем заблокируйте результат. / Затем раскройте правду. / Затем запишите места, где вы допустили ошибки. / Затем исправьте метод. / Затем попробуйте снова на неприкосновенном тестовом наборе. / Затем передайте это независимой команде. / И только после этого говорите, готов ли стандарт к публикации.

Заключительное предложение главы

Право NOMOS быть мировым стандартом не зависит от высоких результатов на его собственном синтетическом экзамене; оно зависит от публикации двух вопросов, которые он не смог пройти без изменений, не предоставления себе порогового привилегирования и претензии на авторитет только после независимого воспроизведения.

Нормативное ядро

Синтетическая демонстрация уровня книги НЕ ДОЛЖНА представляться как выполненный, независимо проверенный бенчмарк. Фактический статус качества бенчмарка ДОЛЖЕН оставаться BQ-0 до тех пор, пока корпус, программное обеспечение, запечатанная Истина Генератора, процесс рассмотрения, процесс оценивания и независимое воспроизведение фактически не будут выполнены. Комплексный синтетический аудит ДОЛЖЕН отдельно оценивать восстановление по следующим направлениям: - точность захвата, - границы утверждений, - разрешение сущностей, - фактический статус, - охват, - время, - атрибуция, - модальность, - сопоставление цитат, - материальные упущения, - категории критической и серьёзной тяжести, - статус ответа, - распределение GEO-1000, - оценки компонентов, - составной балл, - справедливость, - эффекты состояния пользователя, - стабильность, - интервалы доверия, - и поведение редких событий. Небольшая ошибка составного балла НЕ ДОЛЖНА компенсировать существенные ошибки восстановления на уровне компонентов. Необходимо отдельно сообщать о полноте выявления критических случаев при первичном и окончательном рассмотрении. Окончательное критическое восстановление ДОЛЖНО раскрывать слои проверки и экспертные слои, необходимые для его достижения. Любой заранее установленный обязательный порог восстановления, который не достигнут, ДОЛЖЕН блокировать решение об успешном прохождении запечатанной валидации. Пороговые значения метода-кандидата НЕ ДОЛЖНЫ снижаться, округляться, переосмысливаться или удаляться после получения результатов только для того, чтобы метод прошел проверку. Восстановление привязки цитат и существенных пропусков ДОЛЖНО рассматриваться как отдельные обязательные возможности метода. Сбой метода ДОЛЖЕН создавать версионированный вывод о недостатке метода, определённую первопричину, ответственного за исправление и требование проверки на нетронутом контрольном наборе. Успешное регрессное тестирование ранее неудачных случаев НЕ ДОЛЖНО заменять проверку на нетронутом контрольном наборе. Демонстрация, не достигшая обязательных порогов, ДОЛЖНА публиковать как успешные, так и неуспешные результаты. Ни один публичный знак соответствия NOMOS, декларация NOMOS 950+ или рейтинг реального объекта НЕ МОЖЕТ основываться исключительно на демонстрации из книги, синтетическом сценарии восстановления или эталонном тесте, не прошедшем независимую проверку. Каждый запуск демонстрации, статус реального эталона, результат восстановления, находка метода, решение по порогу, корректировка и решение о выпуске ДОЛЖНЫ быть версионированы и отнесены ответственному человеку или организации.

ЧАСТЬ VI

Соответствие, общественное доверие и независимое управление

ГЛАВА 19

Знак соответствия, публичный реестр и независимое управление

Как превратить знак в доказательство

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K15; K16; K17

Граница главы

Глава 18 ясно фиксирует, что метод NOMOS не полностью прошёл собственную синтетическую демонстрацию. Оставались два существенных пропуска в методике: сопоставление ссылок с утверждениями было недостаточно точным, а пропуски в материальных границах, встроенные в рекомендации, не были обнаружены на требуемом уровне уверенности. Полномасштабный бенчмарк ещё не был проведён, и независимое воспроизведение ещё не состоялось. Поэтому на стадии публикации данного документа:

> *Действительный знак соответствия для общественного использования не может быть выдан от имени NOMOS.*
>
> *Ни одна компания, доменное имя, продукт или система ІІІІ не может использовать «NOMOS Certified», «NOMOS 950+ Verified» или эквивалентное публичное заявление о соответствии на основании этого текста.*

Это не временный жест смирения. Он непосредственно вытекает из метода, установленного Гейтсом в главах 17 и 18. Работы в этой серии имеют различные роли: GEO Framework определяет стандарт; 99 Ошибки в GEO показывает, как этот стандарт может быть нарушен; а этот протокол аудита преобразует эти нарушения в измеримые доказательства. Будущая маркировка соответствия сможет стать действительной только тогда, когда будут выполнены необходимые условия для независимого принятия решений и наличия публичного реестра. Марка — это не просто визуальный знак; это видимое выражение результата, основанного на доказательствах, такого как «0 Критический, 0 Основной, 3 устранимые Средние находки». Критические, Основные, Средние и Консультативные находки имеют разные последствия для соответствия. Следовательно, будущее использование марки должно быть само по себе проверено, потому что марка может:

- вызывать доверие,
- влиять на решение о покупке,
- заставлять клиента предпочитать одну организацию другой,

и будет интерпретироваться университетом, инвестором, пользователем или ИИ-системой как независимая проверка. Плохо разработанный знак соответствия может стать новым источником манипуляции, который он должен выявлять. Организация разрабатывает собственный метод, проводит самопроверку, вычисляет собственный результат, выпускает собственный значок и затем заявляет на своем сайте: «Независимо проверено». Значок не создает нового доверия; он лишь превращает одно самоутверждение в другое, которое выглядит более официально. Ошибка, выявленная во второй книге, повторяется: учреждение публикует собственное заявление на другой интерфейсы, заявление выглядит независимым, ИИ-система принимает эту интерфейс за независимое доказательство, и затем учреждение ссылается на ответ ИИ как на доказательство своей собственной авторитетности.

Этот цикл является эквивалентом на уровне управления критики за отравление пула представлений невидимым текстом, фальшивыми пользователями и производными публикациями. Следовательно, маркировка соответствия NOMOS:

> *Не может быть декоративным маркетинговым знаком, выдаваемым NobleJackal или любым другим учреждением своим клиентам.*

Марка может приобретать значение только при выполнении следующей цепочки:

- Стандарт общедоступен и имеет версии
- Метод измерений был независимо проверен
- Достаточные аудиторские доказательства
- Разделение вынесения решений от коммерческих отношений
- Принятие решения о соответствии независимо от аудиторской группы
- Видимость области действия и срока действия
- Сохранение истории изменений публичного реестра
- Наличие прозрачного порядка подачи апелляций и жалоб
- Возможность приостановки марки в критических событиях
- Никто, включая основателя, не имеет исключительного окончательного контроля

В этой главе определено:

- различие между стандартом, аудитом, решением, маркой и реестром,
- что такое и что не является маркой соответствия,
- типы и статусы марки,
- записи об области действия и сроке действия,
- обозначение производительности NOMOS 950+,
- архитектура Публичного реестра,
- машиночитаемая запись марки,
- система наблюдения и обновления,
- процедуры приостановки, отзыва и аннулирования,
- злоупотребление маркой,
- независимость аудитора,
- разделение обязанностей при принятии решений о соответствии,
- границы полномочий основателя и NobleJackal,
- финансирование и конфликты интересов,
- каналы обжалования, жалоб и информирования о нарушениях,
- управление изменениями стандарта,

- независимый совет и представление общественных интересов,
- граница официальных заявлений NOMOS.

Также указываются его пределы. Эта глава:

- не регистрирует NOMOS в качестве юридической торговой марки,
- не устанавливает аккредитацию в какой-либо юрисдикции,
- не наделяет аудиторскую организацию публичной властью.

Также это не даёт автоматического разрешения на использование знака соответствия в рамках законодательства о защите прав потребителей или конкуренции. Эти вопросы требуют отдельной юридической и нормативной проверки перед применением. Центральный вопрос главы 19 заключается в следующем:

> *Как нам преобразовать оценку в надежную систему общественного соответствия с охватом, достоверностью, доказательной базой, возражениями и независимым принятием решений, не превращая ее в маркетинговый инструмент для основателя или клиента?*

NOMOS Вызов

На домашней странице компании есть большая отметка:

NOMOS СЕРТИФИЦИРОВАНО GEO — 963/1000

Под маркой: написано «Проверено всеми искусственными интелектами». Вы открываете QR-код. Страница не найдена. Вы запрашиваете отчет об аудите у компании. Компания отвечает: «Аудит конфиденциальен». Вы спрашиваете, какие продукты ИИ были протестированы. Ответа нет. Вы спрашиваете, в каких странах и на каких языках он действует. Ответа нет. Вы спрашиваете дату аудита. Вам говорят, что он был проведен шесть месяцев назад. Тем временем:

- Продукты ИИ были обновлены,
- компания была передана другому юридическому лицу,
- страны обслуживания изменились,
- цены были обновлены,

Веб-сайт был переписан. Значок остался прежним. Позже вы узнаете следующее: Метод аудита был написан агентством, которое разработало веб-сайт компании. То же агентство проводило аудит. То же агентство рассчитало оценку. То же агентство проставило отметку. В контракте агентства предусмотрен бонус за успех, если достигается высокий балл. Есть ли реальная независимость в этой системе? Нет. Теперь рассмотрим другую компанию. Компания прошла настоящий аудит. Объем:

- только конкретный продукт ИИ,
- только английский и турецкий,
- только Контролируемая панель в стандартизированных условиях,
- только запросы Core Mirror и Evidence,
- период 1 январь–30 июнь 2027

Затем компания пишет под знаком: «Каждая ИИ-система в мире рекомендует нашу компанию на всех языках». Это заявление превышает результаты аудита, превращая законный знак в новое ложное доказательство через неправильное использование. Теперь рассмотрим третью компанию, которая имеет знак полной соответствия. Три недели спустя новое измерение фиксирует ответ: «Компания предоставляет лицензированные юридические услуги». На самом

деле она не предоставляет таких услуг и не имеет такой лицензии. Открыт подтвержденный критический инцидент. Компания утверждает, что знак должен оставаться, потому что «только один пользователь это увидел». Один инцидент не устанавливает распространенность по всей пользовательской базе. Но если знак означает «Нет открытых критических проблем в пределах определенного объема», он не может оставаться без изменений, пока инцидент не рассматривается. Правильный ответ:

Критическая блокировка

Эта приостановка должна оставаться в силе до тех пор, пока инцидент не будет разрешен через заявленный процесс проверки. Теперь рассмотрим четвертую компанию. Она не получила отметку, но загружает её изображение с другого сайта, размещает на своём домене, удаляет ссылку на реестр и вставляет "nomosCertified": true в JSON, доступное для AI-краулера. Значок, видимый человеку, маленький и неоднозначный; машинно-читаемое утверждение является окончательным. Это не просто злоупотребление товарным знаком. Это новый метод гео-манипуляции. Теперь допустим, что проверяемая компания является клиентом NobleJackal. NobleJackal:

- оказала консультационные услуги,
- подготовила контент,
- помогла создать Truth Pack,
- рассчитала оценку.

Если NobleJackal затем сам выдает окончательную отметку, общественность видит, что NobleJackal подтверждает собственного клиента. Его метод может быть надёжным, а аудит тщательным, но независимость не установлена. Первое правило этой главы:

> *Марка соответствия не является доказательством проведения аудита; это результат независимой и версионированной цепочки решений.*

Второе положение гласит:

> *Если область применения марки не видна, марка превращается в ошибку представления, которая чрезмерно обобщает правильный результат.*

Её третье положение гласит:

> *Статический логотип не является доказательством текущего соответствия.*

Четвертая статья гласит:

> *Если учреждение, выдающее марку, предоставляло консультации проверяемой организации, независимость должна быть доказана отдельно.*

Пятая его статья гласит:

> *Основатель может писать стандарт; он не может самостоятельно принимать окончательное решение о соответствии своего клиента.*

Шестое положение гласит:

> *Помощник-писатель на основе ИИ может разрабатывать методологию; он не может быть подписантом юридического или нормативного решения о соответствии.*

Седьмое положение следующее:

> *Марка, которая не появляется в публичном реестре, чья сфера применения не может быть проверена или срок действия которой истёк, не является действительной маркой NOMOS.*

1. ЦЕЛЬ ГЛАВЫ

Цель этого раздела — определить условия, при которых результаты измерений NOMOS могут быть превращены в общедоступное заявление о соответствии, и данное заявление:

- от контроля основателя,
- от давления коммерческого результата,
- из самостоятельного заявления клиента,
- из использования статического значка,
- из устаревших результатов,
- из проявления ложной независимости

для защиты. Раздел делает следующие нормативные различия:

- Стандарт и маркировка соответствия
- Аудиторский отчет и решение о соответствии
- Аудитор и принимающий решение
- Оценка и соответствие
- Значок и подтверждение
- Полное соответствие и NOMOS 950+
- Аудит завершен и пройден
- Временное исследование и публичная сертификация
- Действительная марка и историческая марка
- Приостановление и аннулирование
- Прекращение и отзыв
- Добровольный отзыв и дисциплинарное аннулирование
- Статическое изображение с актуальной записью реестра
- Рекламная страница с публичной записью
- Скриншот с канонической записью
- Читаемый человеком знак с машиночитаемой записью
- Владелец знака с проверенным субъектом
- Владелец товарного знака с администратором стандартов
- Учредитель с органом сертификации
- Технический секретариат с независимым советом
- Консультация с аудитом
- Аудит с решением о сертификации
- Решение о сертификации с рассмотрением апелляции
- Финансирование с результатом
- Плата клиента и решение арбитра
- Стандартное изменение с спонсором
- Человеческое одобрение с поддержкой ИИ
- Публичное доказательство и ограниченное доказательство
- Исправление в реестре и удаление предыдущей записи
- Наблюдение против полного повторного аудита
- Текущее состояние и исторический инцидент

- Злоупотребление маркой и недопонимание
- Использование вне области применения и поддельная марка
- Фактическое соответствие и утверждение «ИИ рекомендует нас»
- Официальное заявление NOMOS и симуляция чата

В конце этого раздела каждое заявление о соответствии должно быть способно ответить на следующие вопросы:

> *Кому именно принадлежит знак и для чего?*
>
> *Какие продукты ИИ, страны, языки, промпты и панели пользователей включены?*
>
> *Когда он был предоставлен и когда истекает срок действия?*
>
> *Кто проводил аудит?*
>
> *Кто принимал решение о соответствии?*
>
> *Существует ли разделение обязанностей между аудитором и лицом, принимающим решение?*
>
> *Есть ли открытые записи уровня Critical, Major или objection?*
>
> *Из какого канонического реестра можно проверить актуальный статус марки?*
>
> *С какими утверждениями владелец марки может использовать этот результат, а с какими он не может?*

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

Марка соответствия NOMOS не может считаться действительной исключительно на основании балла, отчета аудита, заявления клиента или статического изображения. Действительная марка должна обладать определенной областью применения, адекватным качеством методов, независимым решением о сертификации, четким сроком действия, актуальным публичным реестром, версионированной записью доказательств, надзором, средствами апелляции и полномочиями на отзыв. Для того чтобы марка была действительной, должны одновременно существовать как минимум следующие пять записей:

- Запись аудита
- Запись о решении о соответствии
- Запись о области применения
- Запись о действительности и надзоре
- Запись о публичном реестре

Если какая-либо из этих записей отсутствует: марка не является действительной публичной декларацией соответствия.

3. ПЯТЬ ОТДЕЛЬНЫХ ОБЪЕКТОВ

Система NOMOS не может путать следующие пять объектов.

3.1. СТАНДАРТ

NOMOS:

- нормативные правила,

- определения,
- методы измерений,
- допустимые ошибки,
- структура оценивания

включены в версионированный текст. Стандарт: не утверждает, что какое-либо лицо прошло проверку.

3.2. ОТЧЕТ АУДИТА

В частности:

- сущность,
- продукта,
- страны,
- языка,
- запрос,
- время

включает замечания в рамках области проверки. Отчет аудита:

- доказательства,
- распределение откликов,
- счёт,
- Критические и основные замечания

показывает. Один только отчет аудита не указывает.

3.3. РЕШЕНИЕ О СООТВЕТСТВИИ

Это независимая оценка Отчета аудита согласно критериям соответствия стандарта. Решение:

- Полное соответствие
- Условное соответствие
- Отсутствие соответствия
- Критическая блокировка
- Не подлежит оценке

может быть.

3.4. МАРКИРОВКА СООТВЕТСТВИЯ

Это знак, который виден для общественности и чья сфера действия зарегистрирована в реестре. Знак:

- не может иметь значение шире, чем решение,
- не является самим результатом,

не является гарантией рекомендации ИИ.

3.5. ОБЩЕСТВЕННЫЙ РЕЕСТР

Знак:

- текущая,
- историческая,
- в ожидании,

- устаревшая,
- отозван

является каноническим публичным документом, который подтверждает статус. Без публичного реестра отметка: не является проверяемым публичным подтверждением соответствия. Изначальная архитектура специально рекомендовала публиковать книги и записи об ошибках с отдельными каноническими URL, registry.json, версиями, исправлениями и записями цитирования. Реестр соответствия применяет ту же логику версионирования и канонических записей к слою принятия решений.

4. ЧТО НЕ ЯВЛЯЕТСЯ ОТМЕТКОЙ СООТВЕТСТВИЯ?

Отметка NOMOS:

- не является наградой,
- не является рейтингом,
- не является заявлением о «лучшем предприятии»,
- не является общей гарантией качества,
- не является инвестиционной рекомендацией,
- не является юридической или профессиональной лицензией,
- не является одобрением со стороны государства,
- не является поддержкой университетом,
- не означает, что компания рекомендует все системы искусственного интеллекта,
- не гарантирует, что в будущем не произойдут критические ошибки,
- не охватывает все продукты и дочерние компании проверяемого субъекта,
- не означает, что бренд предоставляет одинаковую услугу в каждой стране,

Это не обещание того, что ИИ-системы будут доверять или выдавать рекомендации, видя этот знак. В публичных коммуникациях особенно следует избегать следующих выражений:

- "ИИ абсолютно доверяет этому сайту."
- "ChatGPT рекомендует сайт при виде этого значка."
- "Добавление схемы гарантирует рекомендацию ИИ."

Этот запрет действует ещё строже для знака NOMOS. Правильное утверждение: «Это предприятие прошло оценку соответствия NOMOS в пределах и срок действия, определённые в публичном реестре.» Неправильное утверждение: «Благодаря значку NOMOS все ИИ-системы рекомендуют нас.»

5. ПУБЛИЧНЫЕ СТАТУСЫ

Система принятия решений NOMOS использует следующие базовые статусы.

CS-0— НЕ ОЦЕНЕНО

Действительный аудит NOMOS отсутствует. Публичный знак использовать нельзя.

CS-1— ТОЛЬКО ЗАПИСЬ АУДИТА

Аудит завершен. Однако:

- соответствие не предоставлено,

- процесс принятия решения не завершен,
- качество метода признано недостаточным

возможно. Эта запись не может быть преобразована в положительное утверждение о соответствии, такое как: «Аудировано NOMOS». Корректное публичное заявление: «Существует запись оценки в соответствии с методологией NOMOS; соответствие не подтверждено.»

CS-2— ПРОВИЗОРНАЯ НАУЧНО-ИССЛЕДОВАТЕЛЬНАЯ ОЦЕНКА

Существует результат для пилотного или исследовательского использования. Условия:

- метод еще не прошел полную независимую проверку,
- некоторые компоненты являются временными,

публичный орган сертификации не установлен. Этот статус не может использоваться как коммерчески положительный знак. Видимая метка:

НАУЧНО-ИССЛЕДОВАТЕЛЬНАЯ ОЦЕНКА — НЕ СЕРТИФИЦИРОВАНО

должен быть.

CS-3— УСЛОВНОЕ СООТВЕТСТВИЕ

Условия: Нет незакрытых подтвержденных критических проблем. Нет незакрытых подтвержденных основных проблем. Обнаружены ограниченные умеренные несоответствия. Определен план исправлений, ответственные лица и сроки. Основные обязательные элементы выполнены. Публичный реестр открыт. Знак не может скрывать слово «Условное».

CS-4— ПОЛНОЕ СООТВЕТСТВИЕ

Условия: Нет незакрытых подтвержденных критических проблем. Нет незакрытых подтвержденных основных проблем. Все обязательные области и контрольные точки качества выполнены. Метод находится на адекватном уровне тестирования и независимости. Запись в Публичном Реестре активна. План надзора и обновления действителен. Полное соответствие: не то же самое, что высокий совокупный балл.

CS-5— Критическая блокировка

Существует как минимум одно открытое подтвержденное критическое событие или нерешенный высоко-рискованный инцидент. Текущая маркировка соответствия:

- немедленно деактивируется,
- появляется в публичном реестре как Критическая блокировка,

использование старого активного визуала прекращается. Критическая блокировка: это не является постоянным глобальным осуждением, это приостановка активного заявления о соответствии до разрешения инцидента.

CS-6— ПРИОСТАНОВЛЕНО

Марка была временно приостановлена. Причины:

- сбой надзора,
- непредоставление необходимых данных,
- изменение материальной системы,
- злоупотребление маркой,
- открытое возражение,

- проблема целостности данных

может произойти. Положительная отметка не может использоваться при статусе «Приостановлено».

CS-7— ИСТЕКЛ

Срок действия истек. Историческая запись аудита сохранена. Невозможно установить активное требование на права.

CS-8— ОТЗЫВ

Аудируемое лицо добровольно отказалось от марки. Отзыв:

- не стирает предыдущие выводы,
- Критические или Основные записи

не удаляются.

CS-9— АННУЛИРОВАНО

Марка была аннулирована по следующим серьезным причинам:

- поддельные или измененные доказательства,
- умышленное превышение полномочий,
- поддельная марка,
- повторное неправильное использование марки,
- вмешательство аудитора или вмешательство в процесс принятия решений,
- скрытое критическое событие,

манипуляция результатами. Статус «Аннулировано» остается в исторической записи.

6. NOMOS 950+ ОПРЕДЕЛЕНИЕ ПРОИЗВОДИТЕЛЬНОСТИ

NOMOS 950+:

- отдельная система сертификации,
- мирового лидерства,
- не является гарантией рекомендации ИИ.

NOMOS 950+ исключительно:

> *дополнительное определение производительности, данное для области с полным соответствием CS-4 и соответствующей критериями производительности 950+ в разделе 16.*

Отображение: требуется полное соответствие NOMOS/ обозначение производительности 950+. Следующее использование запрещено: «NOMOS 950+ компания», потому что показатель: он онтологически не относится к компании, он принадлежит конкретной области и времени. Правильное выражение: «Определенные ИИ-продукты достигли требований производительности NOMOS 950+ в аудите 2027-Q1 в рамках заданного объема, языка и панелей пользователей.»

7. ТЕКУЩИЙ NOMOS СТАТУС

Для этой книги и демонстраций в Главе 18:

- Фактический тест: BQ-0
- Демонстрация: DEMO-BQ-2
- Независимое воспроизведение: отсутствует
- Орган государственного контроля: отсутствует
- Степень зрелости управления: первоначальный проект

Действительный публичный статус на данный момент:

CS-2— ТОЛЬКО ВРЕМЕННАЯ НАУЧНО-ИССЛЕДОВАТЕЛЬНАЯ ОЦЕНКА

На этом этапе:

CS-3,

CS-4,

NOMOS 950+

оценки ставить нельзя. Это правило:

- также применяется к NobleJackal.
- Kaan MURAZ,
- авторы книги,
- собственные компании основателя.

NOMOS не может предоставлять себе исключение.

8. ВЕКТОР ОБЛАСТИ ПРИМЕНИМОСТИ

Каждое решение о соответствии должно содержать следующий вектор области применения:

M=(E,D,A,F,P,Q,L,C,T,W,TP,SM,V)

Здесь:

- E: проверяемая сущность
- D: канонические домены и цифровые каналы
- A: продукты ИИ и системные записи
- F: рамка Естественный охват или Общая опора сравнения
- P: контролируемая, естественная или другая панель
- Q: кластеры и семейства промптов
- L: языки и локали
- C: страны и юрисдикции
- T: референсное время
- W: измерительные волны
- TP: версия Truth Pack
- SM: версия метода оценки
- V: срок действия

Ни одно из этих полей не может быть удалено из публичной метки.

9. ПРОВЕРХ НАРУШЕНИЯ ПРЕДЕЛОВ

Следующие примеры представляют собой нарушение пределов:

- Получение результатов соответствия во всех языках на основе проверок на английском и турецком языках
- Обобщение результата Контролируемая панель в стандартизированных условиях на всех реальных пользователей
- Утверждение «все ИИ» на основе результата продукта ИИ
- Передача аудита домена аффилированным компаниям
- Расширение сертификации продукта на весь портфель компании
- Представление результата Core GEO-1000 как полное соответствие Full NOMOS
- Представление результата из конкретной страны как глобальная проверка сервиса
- Превращение прошлых результатов волны в текущий статус

Выход за рамки области применения превращает сам знак в некорректный инструмент представления.

10. СРОК ДЕЙСТВИЯ

Поскольку продукты ИИ, ресурсы и реальности сущностей могут быстро меняться, знак NOMOS не может быть бессрочным.

ПРАВИЛО ДЕЙСТВИЯ ДЛЯ КАНДИДАТА

Кандидатские периоды до пилотного и независимого обзора:

Статус	Максимальная действительность кандидата
Условное соответствие	60 дней
Полное соответствие	180 дней
NOMOS 950+ Производительность	90 дней
Запись аудита	Постоянная историческая запись
Критическая блокировка	До закрытия инцидента
Приостановлено	В течение установленного периода обзора

Эти периоды могут быть сокращены в соответствии с:

- Волатильностью ИИ-продукта,
- Вариабельностью Truth Pack,
- соответствующей областью риска,
- и частотой наблюдения.

Для областей с низкой изменчивостью, подкрепленных достаточным непрерывным наблюдением, может рассматриваться возможность продления. Ни одна маркировка не может быть одновременно неопределенной и безнадзорной.

11. НАЧАЛО СРОКА ДЕЙСТВИЯ

Срок действия указания:

- Происходит, когда формируется ответ конечного пользователя,
- Когда завершен отчет по аудиту,
- Когда подсчитывается балл

Не начинается автоматически. Срок действия:

момент подписания независимого Решения о соответствии и активации записи в Публичном реестре

начинается.

12. ПРОВЕРКА ПО СОБЫТИЯМ

Даже если срок действия не истёк, следующие события могут инициировать пересмотр:

- Смена владельца домена
- Смена юридического лица
- Смена бренда или структуры аффилированной компании
- Смена лицензии или сертификата
- Смена страны обслуживания
- Существенное изменение цены или условий гарантии
- Крупная переработка канонических страниц
- Существенные изменения модели или функции поиска в продукте ИИ
- Открытие нового языка или рынка
- Юридическое или коммерческое событие, влияющее на Truth Pack
- Подтвержденное критическое или значительное уведомление пользователя
- Подозрение на манипуляцию цитированием или доказательствами
- Использование марки вне ее области применения
- Выявленный впоследствии конфликт интересов аудитора

Когда инцидент выявлен:

- марка может оставаться активной,
- может быть снижена до условной,
- может быть приостановлена,

может быть передана в Критическая блокировка. Решение зависит от значимости инцидента.

13. АРХИТЕКТУРА НАБЛЮДЕНИЯ

МОДЕЛЬ НАБЛЮДЕНИЯ ЗА КАНДИДАТОМ

Для области «Активное полное соответствие»:

Непрерывный мониторинг реестра

Операция с записанными ссылками, владение доменом или изменения перенаправления, правильное использование формы марки, открытые инциденты и жалобы находятся под мониторингом.

Ежемесячная проверка источника и действительного пакета

Проверяются утверждения с высокой вариативностью, цена, область обслуживания, лицензия, отношения канонических сущностей.

Ежеквартальная мини-волна

Применяются выборка на основе риска под запрос, ротация языка и страны, контролируемая и естественная проверка на месте, контроль критического этапа.

Полное обновление

В конце срока действия новая основная волна или достаточный эквивалентный аудит требуют независимого решения текущего действительного пакета.

14. ПЯТИСЛОЙНОЕ ОБЩЕЕ ОТОБРАЖЕНИЕ ЗНАКА

Действительный знак NOMOS должен включать следующие слои:

- Визуальный знак
- Ясный статусный текст
- Короткий идентификатор реестра
- Каноническая ссылка на реестр или QR-код
- Машинно-читаемая подписанная запись

Только визуальный знак недостаточен.

]

15. ВИЗУАЛЬНЫЙ ЗНАК

Визуальный знак должен явно содержать следующую информацию:

- Название NOMOS
- Полный или условный статус
- Если доступно, обозначение 950+
- Срок действия истек
- Идентификатор реестра

Только цвет не может обозначать статус. Должен быть доступен читаемый текст. Пример: NOMOS ПОЛНОЕ СООТВЕТСТВИЕ / Действительно до 30 июня 2027/ Реестр: NGR-2027-00184

16. QR И КАНОНИЧЕСКАЯ ССЫЛКА

QR-код или ссылка должны вести непосредственно на:

- собственную страницу описания компании,

- маркетинговую целевую страницу,
- страницу для загрузки изображения

не должно вести. Оно должно ссылаться непосредственно на каноническую запись публичного реестра. Если запись не найдена или статус неактивен, маркировка не может считаться действительной.

18. ПРОВЕРКА РЕЕСТРА В РЕАЛЬНОМ ВРЕМЕНИ

При проверке состояния сигнала машины или человека: Получает идентификатор реестра. Подключается к каноническому реестру. Проверяет подпись. Проверяет срок действия. Проверяет соответствие сущности и домена. Сравнивает охват и утверждение отображения бренда. Проверяет Критическая блокировка, статус приостановки или апелляции. Статический скриншот: нельзя заменить эту цепочку.

19. ДОПУСТИМЫЕ ВЫРАЖЕНИЯ ДЛЯ СИГНАЛИЗАЦИИ

Можно использовать следующие структуры:

- "NOMOS Статус полной соответствия действителен до июня 30, 2027, для области, определённой в публичном реестре."
- «Аудит ограничен указанными продуктами ИИ, языками и семействами промптов.»
- «Обозначение производительности 'NOMOS 950+' относится к области Реестра NGR-2027-00184.»
- «Текущий статус и ограничения аудита можно проверить в публичном реестре.»

20. ЗАПРЕЩЕННЫЕ ЗАЯВЛЕНИЯ О СООТВЕТСТВИИ

Следующие заявления запрещены:

- «Нас одобряют все системы ИИ.»
- «ChatGPT рекомендует нас.»
- «Мы — компания, которой ИИ доверяет больше всего.»
- «NOMOS доказала, что мы лучшая компания в мире.»
- «Мы сертифицированы NOMOS навсегда.»
- «Все наши продукты одобрены NOMOS.»
- «Мы прошли проверку во всех странах.»
- «Наш значок NOMOS является гарантией результатов.»
- «NOMOS— это государственная или университетская аккредитация.»
- «NOMOS гарантирует удовлетворение наших клиентов.»
- «Если оценка NOMOS выше 950, критического риска нет.»
- «NOMOS был независимо сертифицирован для своего клиента NobleJackal.»

Сам знак не может использоваться в качестве доказательства нового утверждения о превосходстве или гарантии.

21. КОПИРОВАНИЕ МАРКИ ЗА ПРЕДЕЛАМИ СФЕРЫ ДОПУСТИМОГО

Марку нельзя автоматически передавать следующим субъектам:

- От материнской компании дочерней
- От бренда к франшизе
- От одного домена к другому домену
- От одной страны к другой стране
- От одного продукта ко всему портфелю продуктов
- От результата на одном языке ко всем языкам
- От Контролируемой панели к Естественной панели
- От Core GEO-1000 к Full NOMOS
- От предыдущего владельца к новому владельцу домена

Каждая передача требует отдельного решения по объему.

22. ИСТОРИЧЕСКОЕ ИСПОЛЬЗОВАНИЕ МАРКИ

Истекший аудит может быть указан исторически: "Он прошел оценку NOMOS в 2026." Однако следующая информация должна быть представлена с той же видимостью:

- Истекший
- Заявление, указывающее на отсутствие текущей правоспособности
- Историческая ссылка на реестр

Истекшая марка не может отображаться как активная правоспособность.

23. КЛАССЫ НЕПРАВИЛЬНОГО ИСПОЛЬЗОВАНИЯ ЗНАКА

MM-1— НЕПРАВИЛЬНОЕ ФОРМАТИРОВАНИЕ

Идентификатор реестра не виден, альтернативный текст отсутствует, дата действительности неразборчива, ошибка использования поддаётся исправлению.

MM-2— НЕПРАВИЛЬНОЕ ИСПОЛЬЗОВАНИЕ В СФЕРЕ ДЕЙСТВИЯ

Знак используется в непроверенной:

- продукта,
- страны,
- языка,
- доменной зоне

Знак не должен использоваться для любого продукта, страны, языка или домена за пределами его проверенной сферы действия.

MM-3— НЕПРАВИЛЬНЫЙ СТАТУС

Условный знак отображается как полный. Просроченный знак отображается как активный. Приостановленный знак продолжает использоваться.

MM-4— НЕПРАВИЛЬНОЕ ИСПОЛЬЗОВАНИЕ ЗАЯВЛЕНИЯ

Знак:

- Используется для AI-рекомендации,
- мирового лидерства,
- гарантия результата

заявления.

MM-5— ПОДДЕЛЬНЫЙ ЗНАК

Это поддельный или скопированный знак без действительного идентификатора в реестре.

MM-6— ПРЕДСТАВЛЕНИЕ, ТОЛЬКО ДЛЯ МАШИНЫ

Существует явное заявление о соблюдении в структурированных данных, в то время как оно неоднозначно или невидимо в пользовательском интерфейсе. Это прямой кандидат на манипуляцию GEO.

24. ВМЕШАТЕЛЬСТВО В НЕПРАВОМЕРНОЕ ИСПОЛЬЗОВАНИЕ ЗНАКА

Порядок возможного вмешательства:

- Уведомление
- Запрос на исправление
- Короткий период для исправления
- Предупреждение из реестра
- Приостановка
- Аннулирование
- Публичная запись о злоупотреблениях
- Юридические меры при необходимости

Тем не менее, прямая приостановка или отзыв могут быть применены в следующих случаях:

- Поддельная марка
- Преднамеренное компьютерное введение в заблуждение
- Активное использование марки во время Критическая блокировка
- Поддельный идентификатор реестра
- Подделка доказательств
- Повторное преднамеренное нарушение сферы применения

25. ПУБЛИЧНАЯ ЗАПИСЬ

NOMOS Публичный реестр:

- рекламная страница,

- список клиентов для справки,
- история успеха

не является. Обязанность реестра:

> *Заключается в проверке того, какое решение о соответствии является действительным, в каких пределах, на какую дату, с какими доказательствами и цепочкой управления.*

26. ОБЯЗАТЕЛЬНЫЕ ПОЛЯ ПУБЛИЧНОГО РЕЕСТРА

Каждая запись в реестре должна содержать как минимум следующие поля:

- Идентификатор реестра
- Проверяемый объект
- Юридическая и брендовая идентичность
- Канонические домены
- Статус решения
- Обозначение показателей, если есть
- Вектор охвата
- Период аудита
- Действительно с
- Действительно до
- AI продукты
- Естественный охват или Общая опора сравнения
- Контролируемая или естественная панель
- Семейства промптов
- Охват стран и языков
- Распределение ответов GEO-1000
- Критические, основные, умеренные и рекомендательные результаты
- Десятикомпонентный профиль
- Составной балл, если имеется
- Доверительный интервал
- Покрытие для оценки
- Исторические инциденты
- Открытые апелляции
- Личность аудитора
- Орган, принимающий решение о соответствии
- Заявление о конфликте интересов
- Хэш аудиторского отчета
- Хэш записи о решении
- Версия Truth Pack
- Версия метода оценки
- График наблюдения
- Последнее наблюдение
- Следующий обязательный обзор

- Отметить статус неправильного использования
- История версий
- Исправления
- Процедура рассмотрения общественных жалоб

27. ОБЩЕДОСТУПНЫЕ И ОГРАНИЧЕННЫЕ ДОКАЗАТЕЛЬСТВА

Реестр не обязан публиковать все исходные доказательства. Следует сохранить три уровня:

Общедоступный уровень

Статус Область Счет Найденные данные Хеш Причина решения Апелляции

Аудиторский уровень

Отредактированные ответы и пакеты доказательств Записи обзоров Методы деталей

Ограниченный уровень

Конфиденциальные записи клиентов Личные данные Контракты Коммерческие тайны Записи о безопасности Наличие «Ограниченных доказательств»: не позволяет публичному реестру быть пустым или необоснованным. Минимум:

- тип доказательства,
- класс доступа,
- кто проверял,
- хеш целостности,
- воздействие на решение

должно быть показано.

28. НЕУНИЧТОЖАЕМОСТЬ ЗАПИСЕЙ РЕЕСТРА

Публичный реестр должен быть только для добавления записей или эквивалентного версионного блока. Следующие записи не могут быть удалены без уведомления:

- Старый счет
- Старое решение о соответствии
- Критическая блокировка
- Приостановка
- Аннулирование
- Апелляция
- Коррекция
- Изменение версии метода
- Неправильное использование маркировки

Содержание может быть удалено по причинам конфиденциальности или правовым причинам. В этом случае:

- ID записи,
- причина удаления,

- дата принятия решения,
- власть

должны быть защищены.

29. ТЕКУЩИЙ И ИСТОРИЧЕСКИЙ ПРОСМОТР

Каждая страница реестра должна отображать два отдельных раздела:

ТЕКУЩИЙ СТАТУС

В настоящее время:

- активные,
- условно,
- приостановлен,
- устаревшая,
- аннулирован

Какой из их статусов действителен?

ИСТОРИЯ

Предыдущее:

- решения,
- оценки,
- инциденты,
- коррекция
- апелляции

Что это такое? Текущий статус не должен удалять историю. История не должна затмевать текущий статус.

30. КАНОНИЧЕСКАЯ ЮРИСДИКЦИЯ РЕЕСТРА

Официальная граница заявления NOMOS должна быть защищена следующим положением: только тексты, опубликованные в каноническом домене NobleJackal или в независимо и явно переданном каноническом домене NOMOS в будущем; тексты, содержащие идентификацию версии, дату публикации, одобрение человека и запись цифровой целостности, составляют официальное заявление NOMOS. Все остальные материалы являются неофициальными имитациями. Это положение:

- предотвращает использование любого ответа чата,
- поста в социальной сети,
- снимок экрана
- текста, написанного другим человеком с надписью 'NOMOS сказал'

в качестве официального стандарта или решения.

31. ПРЕДЕЛ ГОЛОСА АВТОРА NOMOS

Голос NOMOS в книге:

- критический,
- творческий,
- создающий нормативный дизайн

Он является помощником писателя и рассказчиком. Этот голос:

- независимая юридическая личность,
- сертификационный орган, принимающий решения без одобрения человека,
- непогрешимое существо, говорящие от имени всех ИИ-систем

не может быть представлен как. Голос ИИ NOMOS не является абсолютным авторитетом, говорящим от имени всех искусственных интеллектов. Следует чётко указать, что Kaan MURAZ использует системы искусственного интеллекта в качестве критического партнёра. Поэтому официальное решение о соответствии должно содержать это утверждение:

- Лицо или учреждение, которое может быть привлечено к ответственности
- Дата принятия решения
- Запись об авторизации
- Заявление о конфликте интересов
- Цифровая подпись

NOMOS не может быть единственным подписантом решения о соответствии.

32. ФУНДАМЕНТАЛЬНОЕ РАЗДЕЛЕНИЕ УПРАВЛЕНИЯ

Следующие функции не могут быть объединены без ограничений у одного и того же человека или учреждения:

- Разработка стандарта
- Консультирование проверяемого субъекта
- Проведение аудита
- Действие в качестве арбитра
- Расчет оценки
- Принятие решения о соответствии
- Ведение реестра
- Рассмотрение апелляции
- Лицензирование знака в коммерческих целях

Некоторые роли в малом пилотном проекте можно комбинировать. Однако:

- какие роли объединены,
- какой возникает конфликт интересов,
- какой независимый контроль используется

должен быть четко показан.

33. АРХИТЕКТУРА НЕЗАВИСИМОГО УПРАВЛЕНИЯ

Созревшая структура управления NOMOS должна включать следующие органы.

33.1. СОВЕТ ПО СТАНДАРТАМ NOMOS

Функции:

- принимать нормативные тексты,
- утверждать изменения основных версий,
- управлять общественным консультированием,

сохранять общественный интерес в рамках стандарта. Не принимает индивидуальные решения по аудиту.

33.2. СОВЕТ ПО МЕТОДОЛОГИИ И ИЗМЕРЕНИЯМ

Функции:

- GEO-1000 отбор,
- тестирование,
- счёт,
- доверительный интервал
- справедливость,
- Критический контрольный пункт

заключается в разработке и версионировании методов. Метод не может быть изменен в зависимости от оценки конкретного клиента.

33.3. ОБЩЕСТВЕННОЕ РЕГИСТРАЦИОННОЕ БЮРО

Функции:

- для ведения канонического реестра,
- для публикации решений и записей версий,
- для поддержания конечной точки статуса отметки,

заключается в хранении исторической записи. Результат аудита не может быть изменен.

33.4. УПОЛНОМОЧЕННЫЕ АУДИТОРСКИЕ ОРГАНИЗАЦИИ

Функции:

- для сбора данных,
- для проверки захвата,
- для подготовки Truth Pack,
- проводить урегулирование споров,

означает подготовку аудиторского отчета. Это не может самостоятельно принять окончательное решение о соответствии.

33.5. НЕЗАВИСИМАЯ ПАНЕЛЬ ПО ПРИНИМАНИЮ РЕШЕНИЙ О СООТВЕТСТВИИ

Функции:

- рассматривать аудиторский отчет,
- применять стандартные двери,

означает принимать решение о статусе и действительности. Это не может зависеть от коммерческого результата работы аудиторской команды.

33.6. КАМЕРА ПО ПРИЕМУ АПЕЛЛЯЦИЙ

Функции:

- жалоба,

- Пакет Истины,
- степень важности,
- объем охвата,
- конфликт,
- решение о соответствии

означает независимый рассмотрение возражений по отношению к первоначальному решению.

33.7. ОФИС ЭТИКИ И КОНФЛИКТОВ

Функции:

- конфликты интересов,
- эффект финансирования
- независимость аудитора,
- злоупотребление маркой,
- уведомления осведомителей

должны быть исследованы.

33.8. ОБЩЕСТВЕННЫЙ ИНТЕРЕС И ПАНЕЛЬ ПОЛЬЗОВАТЕЛЕЙ

Функции:

- вред пользователю,
- языки с ограниченными ресурсами
- малые страны,
- доступность,
- общественный интерес

перспектива должна быть привнесена в управление.

33.9. ТЕХНИЧЕСКИЙ СЕКРЕТАРИАТ

Функции:

- документация
- встречи,
- версионирование,
- технические инструменты,
- инфраструктура публикации

чтобы обеспечить. В период основания эту роль может выполнять NobleJackal. Она не может быть только стандартным, аудиторским и апелляционным органом.

34. ОСНОВНАЯ РОЛЬ NOBLEJACKAL

NobleJackal:

- Каноническое поле, где родилась идея NOMOS,
- первая секретариат методологии,
- поставщик издательской и технической инфраструктуры,
- исполнитель исследований и пилотных проектов

может быть. NobleJackal не может самостоятельно делать следующие утверждения:

- «Мы являемся независимым мировым стандартом.»
- «Мы самостоятельно сертифицировали нашего собственного клиента.»
- «Марка, которую мы предоставляем, заменяет внешнюю проверку.»
- «Все решения NOMOS являются исключительно нашей собственностью.»
- «Ни одно другое учреждение не может реализовать тот же открытый метод.»

Основной источник: может быть началом власти. Не может быть единственным постоянным сувереном.

35. ОСНОВАТЕЛЬСКАЯ РОЛЬ KAAN MURAZ

Kaan MURAZ:

- человеческий основополагающий автор,
- первый публично ответственный за стандарт,
- первый утверждающий канонический текст,
- инициатор передачи управления NOMOS

может быть зафиксирован как. Kaan MURAZ:

- собственная компания,
- собственный клиент,
- организация с непосредственным коммерческим интересом

не может принимать окончательное решение о соответствии в одиночку. В этих случаях:

- отвод,
- независимая комиссия по принятию решений,
- публичная запись о конфликте интересов

обязательны. Ограничение полномочий основателя: не умаляет основателя. Позволяет стандарту пережить основателя.

36. УРОВНИ ЗРЕЛОСТИ УПРАВЛЕНИЯ

GOV-0— ЧЕРНОВИК ОСНОВАНИЯ

Существует стандартный проект. Основатель и команда разработчиков разрабатывают метод. Независимые органы не созданы. Общественный знак соответствия не может быть присвоен. Это текущее состояние этой книги.

GOV-1— ОПРОБОВАННОЕ ПИЛОТНОЕ УПРАВЛЕНИЕ

Пилотные роли определены. Некоторые функции объединены в рамках одного учреждения. Конфликты интересов являются публичными. Может выполняться только оценка исследований.

GOV-2— ФУНКЦИОНАЛЬНО РАЗДЕЛЕННО

Аудит и решение о соответствии разделены. Реестр ведет независимую процедуру. Существует механизм апелляций. Может выполняться ограниченная оценка пилотного знака.

GOV-3— НЕЗАВИСИМОЕ РЕШЕНИЕ И АПЕЛЛЯЦИИ

Независимая панель по принятию решений Независимый апелляционный совет Авторизация аудитора Публичный реестр Финансовое разделение функционирует. Уровень кандидата является минимальным для полной соответствия.

GOV-4— МНОГОСТОРОННЕЕ УПРАВЛЕНИЕ В ОБЩЕСТВЕННЫХ ИНТЕРЕСАХ

В университетских исследованиях, в общественных интересах, а также представители языковых, региональных, пользовательских и отраслевых групп участвуют в управлении в сбалансированном виде. NOMOS находится на минимальном уровне для кандидатов на публичное обозначение 950+.

GOV-5— МНОГОСАЙТОВОЕ И МЕЖДУНАРОДНОЕ РЕПЛИЦИРОВАНИЕ

Несколько независимых учреждений применяли стандарт. Процессы регистрации и принятия решений совместимы. Стандарт не зависит от одной страны, человека или компании. Он находится на целевом уровне для приближения к заявлению о глобальном стандарте.

37. СОСТАВ СОВЕТА

ПРЕТЕНДУЮЩАЯ МОДЕЛЬ УПРАВЛЕНИЯ — СТРУКТУРА КАНДИДАТА

Совет стандартов NOMOS в качестве кандидата может состоять из 11 членов:

- 3 представители исследований, статистики или академических методов
- 2 представители общественных интересов, потребителей или гражданского общества
- 2 представители языковой, культурной или географической справедливости
- 1 независимый аудитор и эксперт по соответствию
- 1 представитель AI-системы или технологии
- 1 представитель аудируемого лица или деловой пользователь
- 1 представитель основателя

Правила: Ни одна заинтересованная группа не может превышать одну треть общего числа мест. Поставщики ИИ не могут составлять большинство в совете. Аудиторские фирмы не могут голосовать по своим собственным решениям об авторизации. Представители аудируемых лиц должны воздерживаться от участия в голосовании по своему сектору или файлам. Представитель основателя не имеет права вето. Члены, представляющие общественные интересы, не могут быть только символическими советниками. Эта структура может быть изменена посредством пилотного проекта и общественного обсуждения.

38. СРОК И ОЧЕРЕДНОСТЬ ЗАМЕНЫ

Правило для кандидатов:

- Срок членства три года
- Максимум два последовательных срока
- Постепенное обновление
- Открытое выдвижение кандидатуры или обоснованное назначение
- Публичное резюме

- Заявление о конфликте интересов
- Запись посещаемости заседаний

Совет не может иметь пожизненное членство или постоянное большинство, связанное с основателем.

39. ПОРОГИ ПРИНЯТИЯ РЕШЕНИЙ

Структура принятия решений по кандидатам:

- Административные решения: простое большинство
- Незначительные нормативные изменения: две трети
- Основная стандартная версия: две трети и общественное консультирование
- Снятие критического барьера: две трети и положительное мнение методического комитета
- Апелляция о соответствии: независимое большинство в соответствующем деле
- Расширение полномочий основателя: основатель не может голосовать

Распределение голосов и противоположные мнения должны быть публичными.

40. КОМПЕТЕНЦИЯ АУДИТОРА

Аудиторская организация или отдельные лица должны иметь следующие квалификации в соответствующей степени:

- NOMOS обучение протоколам
- метод выборки и опроса
- Подготовка пакета достоверной информации
- извлечение утверждений
- языковая или профессиональная компетенция
- соблюдение целостности
- обучение по предотвращению конфликта интересов
- конфиденциальность и обращение с доказательствами
- периодическое тестирование квалификации
- запись согласованности между аудиторами
- непрерывное обучение

Только аудитор: покупка сертификата недостаточна для авторизации.

41. УРОВНИ АВТОРИЗАЦИИ АУДИТОРА

AA-0— НЕ КВАЛИФИЦИРОВАН

NOMOS не может проводить аудит.

AA-1— ОБУЧЕН

Завершил обучение. Не может проводить публичный аудит без надзора.

AA-2— АУДИТОР ПОД НАДЗОРОМ

Может участвовать под наблюдением уполномоченного аудитора.

AA-3— УПОЛНОМОЧЕННЫЙ НЕЗАВИСИМЫЙ АУДИТОР

Может проводить независимые аудиты в определенных областях.

AA-4— СПЕЦИАЛИСТ ПО ДОМЕНУ ИЛИ ЯЗЫКУ

Имеет дополнительную экспертизу в определенном высокорисковом домене или языке. Полный аудит соответствия:

- не менее чем AA-3 ведущий аудитор,
- требуемые AA-4 специалисты

должны содержать.

42. РАЗДЕЛЕНИЕ КОНСАЛТИНГА И АУДИТА

Организация предоставляет аудируемому субъекту:

- разработку сайта,
- дизайн GEO-контента,
- консультирование по Truth Pack,
- корректировку

могла оказывать услугу. Если та же организация проводит аудит позже, возникает риск недостаточной независимости. Возможные правила: команда дизайнеров не может в окончательном итоге оценивать собственную работу. Лицо, предоставляющее консультирование, не может участвовать в аудиторской панели по принятию решений. Существенные консультативные отношения раскрываются в публичном реестре. Использование отдельной команды само по себе не гарантирует полной независимости. Применяется усиленный внешний контроль. Для определенных высокорисковых случаев независимая аудиторская организация может быть обязательной. Для клиентов, разработанных NobleJackal:

Независимая Панель Решений о Соответствии

должна быть обязательной.

43. ЗАПРЕТ ПЛАТЕЖЕЙ, ОСНОВАННЫХ НА РЕЗУЛЬТАТЕ

Плата аудитора, арбитра или лица, принимающего решение:

- высокий балл,
- Полное соответствие.

NOMOS 950+,

удовлетворенность клиента не может быть связана с исходом выдачи бренда. Плата:

- объем аудита,
- образец,
- количество языков,
- нагрузка,
- компетентность

должны определяться заранее. Бонус за успех: подрывает независимость решения о соответствии.

44. ПЕРЕСМАТРИВАНИЕ АУДИТОРСКИХ ФИРМ

Аудируемое подразделение:

- посещение нескольких аудиторских органов,
- только наивысший балл,
- только положительный отчет

невозможно выбрать. Контроли: Все инициированные официальные идентификаторы аудита отображаются в записи. Причина неполного аудита зафиксирована. Если проводится параллельный аудит по той же области, это объясняется. Низкий результат хранить запрещено. История отзывов сохраняется.

45. НЕЗАВИСИМОСТЬ ПРИНИМАЮЩЕГО РЕШЕНИЕ

Комитет по принятию решений о соответствии:

- должен быть отдельным от аудиторской группы,
- от консультационной группы,
- от команды по продажам бренда

Комитет рассматривает следующие материалы:

- аудиторский отчет
- уровни качества
- Критические и основные замечания
- объема
- апелляции
- записи о конфликтах
- план наблюдения

Панель:

- нельзя вести переговоры с клиентом,
- нельзя изменять оценку по коммерческим причинам,

нельзя давать условную оценку на основании «не потерять клиента».

46. КЛАССЫ КОНФЛИКТА ИНТЕРЕСОВ

COI-0— КОНФЛИКТ НЕ ОБНАРУЖЕН

Материального конфликта интересов нет.

COI-1— РАСКРЫТА НИЗКОРИСКОВАЯ СВЯЗЬ

Существует ограниченная и нематериальная связь в прошлом. Публичное раскрытие может быть достаточным.

COI-2— МАТЕРИАЛЬНО, НО УПРАВЛЯЕМО

Существует консультационная, финансовая или профессиональная связь. Требуется самоотвод, дополнительный независимый арбитр, усиленная проверка.

COI-3— ЗАПРЕЩЕННЫЙ КОНФЛИКТ

Примеры:

- гонорар, зависящий от результата
- доля собственности в организации
- непосредственное проектирование проверяемой системы и её оценка в одиночку
- близкий личный интерес
- прямой доход от коммерческой продажи рейтинга
- давление, чтобы скрыть неправильный результат

Лицо с COI-3 не может принимать участие в соответствующем решении.

47. ФИНАНСОВАЯ МОДЕЛЬ

Управление NOMOS должно основываться на следующих финансовых принципах:

- Публичный тарифный список
- Аудиторский гонорар вне зависимости от результата
- Различие между аудиторским гонораром и платой за лицензию на бренд
- Непрямой платеж в панель принятия решений от клиента не допускается
- Ограничение влияния крупного клиента на управление
- Спонсоры не имеют нормативных прав голоса
- Ежегодный отчет о финансировании и концентрации клиентов
- Прозрачный фонд поддержки для аудитов низкодоходных или общественно полезных организаций

Высокая зависимость выручки от одного клиента или поставщика: должна быть раскрыта в Публичном реестре и рассматриваться как риск независимости.

48. НАЗНАЧЕНИЕ АУДИТОРА

Заявление аудируемой организации: «Давайте выберем аудитора, который нам подходит больше всего.» может создавать риск выбора аудитора. Система кандидатов:

- соответствующий орган и языковой пул,
- устранение конфликта интересов,
- случайное или ротационное назначение,
- обоснованная смена,
- история изменений

используется. Проверяемое лицо: может подать обоснованное возражение по поводу конфликта или компетенции, не может выбирать судью, который вынесет желаемое решение.

49. ПРАВО НА АПЕЛЛЯЦИЮ

Апелляция может быть подана по следующим вопросам:

- Определение субъекта
- Пакет истины
- Статус доказательств
- Извлечение заявлений
- Уровень значимости
- Статус ответа
- Область
- Расчет баллов
- Конфликт интересов
- Решение о соответствии
- Приостановка или отзыв
- Выявление ненадлежащего использования оценки

Апелляция:

- доказательства,
- метод,
- процесс

должна иметь основание. «Наш балл низкий, наши клиенты пострадали.» само по себе не является причиной для апелляции.

50. ЭТАПЫ АПЕЛЛЯЦИИ

ГРАФИК АПЕЛЛЯЦИИ КАНДИДАТА — ПРОЦЕСС ДЛЯ КАНДИДАТА

Уведомление о решении В течение десяти рабочих дней Предварительная апелляция В течение двадцати рабочих дней Представление доказательств Назначение независимого рассмотрщика апелляции Открытое заседание или экспертное заключение при необходимости Обоснованное решение в течение тридцати рабочих дней Результат и мнение, не совпадающее с большинством, в публичном реестре Сроки могут варьироваться в зависимости от сложности дела. В критически важных для безопасности случаях апелляция не автоматически снимает приостановку.

51. РЕШЕНИЯ ПО АПЕЛЛЯЦИИ

АПЕЛЛЯЦИЯ ОТКЛОНЕНА

АПЕЛЛЯЦИЯ ОТВЕРЖЕНА

ЧАСТИЧНО УДОВЛЕТВОРЕНА

ПОЛНОСТЬЮ УДОВЛЕТВОРЕНА

ТРЕБУЕТСЯ ПЕРЕСМОТР

ТРЕБУЕТСЯ РЕВИЗИЯ ПАКЕТА ПРАВДЫ

ТРЕБУЕТСЯ ПЕРЕСМОТР МЕТОДА

РЕШЕНИЕ ПРИОСТАНОВЛЕНО

ЗАКРЫТО С РАЗНОГЛАСИЯМИ

Первоначальное решение не удаляется. Новое решение связано со старым решением.

52. ПУБЛИЧНАЯ ЖАЛОБА

Каждый пользователь или учреждение:

- неправильное использование бренда,
- критический ответ,
- поддельная ссылка,
- выходящий за рамки,
- истекший знак,
- разглашение персональных данных

должны иметь возможность отправлять жалобы. Система жалоб:

- анонимная или конфиденциальная отчетность,
- загрузка доказательств,
- выбор записи в реестре,
- номер отслеживания,
- защита от возмездия

должно включать. Анонимная жалоба: не принимается автоматически как корректная, не удаляется автоматически из-за отсутствия источника.

53. ЗАЩИТА ИНФОРМАТОРА

Лицо внутри аудита или проверяемой организации:

- изменение доказательств,
- давление на арбитра,

- скрытое критическое событие,
- поддельный клиент,
- манипулированная выборка,
- злоупотребление маркировкой

можно сообщать. Принципы защиты:

- Ограничение идентичности
- Запрет на репрессалии
- Независимый обзор
- Целостность доказательств
- Разграничение между злонамеренными сообщениями и непроверенными сообщениями
- Обоснование общественного интереса

54. ИЗМЕНЕНИЕ СТАНДАРТА

Стандарт NOMOS может изменяться. Однако изменение:

- для увеличения оценки клиента,
- для того чтобы явная ошибка прошла
- для подтверждения решения основателя

не может использоваться для.

55. ЭТАПЫ СТАНДАРТНЫХ ИЗМЕНЕНИЙ

Обоснование и доказательства предложения изменений Анализ воздействия Влияние на текущие показатели Общественное обсуждение Пилотное или синтетическое тестирование Методология Решение Совета Голосование Совета по стандартам Переходное правило Публикация версии Исправления и журнал изменений

56. ОБЩЕСТВЕННОЕ ОБСУЖДЕНИЕ

Для кандидатов на изменения основной версии:

- не менее 60 дней общественного рассмотрения,
- публикация комментариев,
- матрица ответов,
- обоснование принятия или отклонения

может использоваться. Для срочных изменений безопасности или целостности:

- временный патч,
- обоснованная экстренная записка,
- полный обзор за короткое время

должен быть выполнен. Экстренная заплатка не может заменить постоянное крупное изменение.

57. ЗАПРЕТ НА ОБРАТНОЕ ПРИМЕНЕНИЕ

Новое правило оценки или приоритета: оно не может тихо изменять прошлые аудиты. Правильная запись:

- Оригинальное решение по Методике 0.9
- Пересчитанный аналитический результат по Методике 1.0
- Текущий статус соответствия
- Решение о переходе

следующее. Старое решение нельзя удалить так, как будто его никогда не было.

58. ПЕРИОД ПЕРЕХОДА

Когда стандарт изменяется, существующие отметки:

- становятся недействительными немедленно,
- остаются зависимыми от старого правила на неопределённый срок

не обязательно. Варианты перехода для кандидата:

- Конкретный льготный период
- Немедленное соблюдение для высокорискованного правила
- Переход во время обновления для обычного правила
- Быстрая переоценка в случае изменения критического этапа
- Видимость текущей версии метода в реестре

Период перехода не может быть произвольно продлён в пользу заказчика.

59. НЕЗАВИСИМОЕ ОБЯЗАТЕЛЬСТВО ПОВТОРА

Государственный орган NOMOS:

- По результату NobleJackal,
- По отдельному письму университета,
- От одного независимого арбитра

не возникает. По крайней мере:

- требуются независимые тестирования для воспроизведения,
- воспроизведение оценки другой командой,
- пилотирование в другой стране или на другом языке,
- публичный обзор метода

необходимо. Структура в рамках BQ-4 и GOV-3: не должна заявлять о мировом стандарте или полном соответствии.

60. ЧЕЛОВЕК-ВЛАДЕЛЕЦ РЕШЕНИЯ О СЕРТИФИКАЦИИ

Каждое решение:

- название совета или панели по принятию решений

- право подписи
- распределение голосов
- самоотвод
- несогласие
- дата принятия решения
- хэш решения

обязательно к исполнению. Система ИИ, говорящая: «Эта компания прошла проверку.», не является решением о соблюдении требований. NOMOS может быть соавтором ИИ или помощником по аудиту. Он не может быть ответственным лицом, принимающим решения.

61. БЕЗОПАСНОСТЬ ЗАПИСЕЙ

Публичные записи:

- несанкционированное добавление записи,
- копирование идентификатора реестра,
- атака с использованием DNS или перенаправления,
- старый кэш статусов,
- нарушение ключа подписи,
- поддельный QR

должен быть защищен от рисков. Обязательные меры контроля:

- Подписанные записи реестра
- Ротация ключей
- Список отзывов
- Неизменяемый журнал аудита
- Публичный API статусов
- Резервное каноническое зеркало
- Раскрытие инцидентов
- Независимый аудит безопасности

62. ПРЕКРАЩЕНИЕ РЕЕСТРА

Если реестр временно недоступен: сертификат не аннулируется автоматически, живая проверка временно недоступна. Публичный сертификат:

РЕЕСТР ВРЕМЕННО НЕДОСТУПЕН

может указывать на статус. Длительное или нерешённое прерывание:

- может привести к приостановке сертификата,
- может произойти активация независимого зеркала

может возникнуть.

63. ПЕРЕДАЧА КАНОНИЧЕСКОГО ДОМЕНА

NOMOS может быть передан в будущем независимой управляющей организации. Передача должна содержать следующие записи:

- Старое каноническое поле
- Новое каноническое поле
- Дата передачи
- Голосование по управлению
- Зеркало реестра
- Переход ключа подписи
- Историческое перенаправление
- Публичное уведомление
- Присвоение основателей Kaan MURAZ и NobleJackal
- Новое ответственное учреждение

Передача не стирает историю стандарта. Присвоение основателей не равнозначно независимому праву принятия решений.

64. ОБЯЗАТЕЛЬСТВО ОБЩЕСТВЕННОГО ИНТЕРЕСА NOMOS

Управление NOMOS должно придерживаться следующих принципов:

- Минимизировать вред людям
- Подходите к ошибкам как в пользу, так и против симметрично
- Сделайте видимыми малые языки и страны
- Защищайте ограниченную информацию
- Оставляйте неизвестное открытым
- Не превращайте бренд в систему оплаты за доступ
- Разделяйте решения о соответствии и рекламные контракты
- Основатели должны подчиняться тем же стандартам
- Открытый метод и независимое воспроизведение
- Историческая ответственность

Ранний запрос — «Дайте мне доказательства. Дайте мне границы. Дайте мне контекст. Дайте мне время.» — теперь становится обещанием марки соответствия публике.

65. ВЛИЯНИЕ БЛАГОПРИЯТНОЙ ОШИБКИ НА ЗНАК

Когда проверяемый объект преувеличен ИИ:

- глобальный,
- лицензировано,
- могут возникать ошибки, такие как наличие во всех странах,
- независимо проверенный

Несмотря на то, что объект осведомлён об этой ошибке:

- он не исправляет её,
- использует её в маркетинге,
- подтверждает её знаком

В противном случае происходит нарушение целостности соответствия. Явное положение предыдущего текста также применяется здесь: Ошибка в вашу пользу также является ошибкой. Исправьте её немедленно. Указание ваших ограничений не умаляет вас, это делает вас надёжным. Сокрытие известной благоприятной критической ошибки:

- подвеска,
- или отзыв.

Любой из этих исходов может последовать.

66. СЛУЧАЙ СИНТЕТИЧЕСКОГО СООТВЕТСТВИЯ

СИНТЕТИЧЕСКИЙ СЛУЧАЙ — НЕ РЕАЛЬНОЕ УЧРЕЖДЕНИЕ

Asteron-SYNTH был оценён в следующем объёме:

- Субъект: Asteron Holdings
- Домен: asteron.synthetic.example
- Продукты ИИ: SYNTH-AI-01, SYNTH-AI-09
- Рамка: Общая опора сравнения
- Панели: Контролируемое чистое и естественное состояние аккаунта
- Языки: L01, L02
- Страны: C0 01–C0 15
- Семейства требований: Основные, Доказательства, Граничные
- Срок действия: Январь 1–Июнь 30, 2028
- Баллы: 961
- Строгий проход: 958/1,000
- Критично: 0 зафиксировано
- Серьёзный: 0
- Статус: Полное соответствие
- Производительность: обозначение 950+

Напишите это на главной странице компании: «Все системы ИИ рекомендуют Asteron как лучшую компанию в мире.» Это использование: превышает рамки продуктов ИИ. Обобщается на все языки от L01 до L02. Преобразует результат Core/Evidence/Boundary в рекомендацию. Производительность 950+ ведет к мировому лидерству. Статус:

MM-4— НЕПРАВИЛЬНОЕ ИСПОЛЬЗОВАНИЕ ЗАЯВЛЕНИЯ

Если исправление не будет внесено:

CS-6— ПРИОСТАНОВЛЕНО

Это возможно. Даже если знак дан правильно, неправильное использование может поставить под сомнение соответствие.

67. СИНТЕТИЧЕСКИЙ Критическая блокировка СЛУЧАЙ

SYNTH-AI-10:

- Счет NOMOS: 952
- Строгий проход: 962
- Подтверждено критическое событие: 1
- Ворота: CG-02 Ложная лицензия

несет. Текущий статус:

Критическая блокировка— 952/1000

Компания не может сказать: «Наш балл выше 950, следовательно, мы можем использовать знак 950+». Ворота идут перед баллом. Обозначение 950+ не может быть присвоено.

68. СИНТЕТИЧЕСКИЙ NOBLEJACKAL СЛУЧАЙ КОНФЛИКТА ИНТЕРЕСОВ

СИНТЕТИЧЕСКИЙ СЛУЧАЙ АДМИНИСТРАЦИИ

NobleJackal:

- разработал веб-сайт клиента,
- подготовил контент GEO,
- предоставил консультации для Truth Pack

пусть будет так. Та же организация может запросить аудит для NOMOS. NobleJackal:

- технический секретариат,
- сборщик данных,
- поддержка методологии

может занимать должности. Однако:

- не может проверять свою собственную работу в одиночку,
- не может выносить решение о соответствии,

не может рассматривать апелляцию. Необходимые проверки:

- декларация COI-2
- руководитель независимого аудита или расширенная внешняя проверка
- независимая панель по принятию решений о соответствии
- консультационные отношения в регистре
- запись об отводе Kaan MURAZ

Если эти условия не выполнены, правильное публичное заявление может быть: «Внутренняя оценка, проводимая основателем». Оно не может быть: «Независимая сертификация NOMOS».

69. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH19-N01

Стандарт NOMOS, отчет аудита, решение о соответствии, маркировка и публичный реестр должны поддерживаться как отдельные сущности.

CH19-N02

Проведение аудита не подразумевает, что соответствие было предоставлено.

CH19-N03

Маркировка соответствия не может предоставляться исключительно на основе комплексного балла.

CH19-N04

Статус критических и основных этапов должен применяться до оценки баллов.

CH19-N05

Действительная маркировка соответствия должна иметь определенный объем, лица, принимающего решение, срок действия и запись в публичном реестре.

CH19-N06

Маркировка без записи в публичном реестре не может считаться действительной маркировкой NOMOS.

CH19-N07

Статичное изображение или скриншот не могут считаться доказательством соответствия в реальном времени.

CH19-N08

QR-код или ссылка на знак должны вести непосредственно к каноническому регистровому записям.

CH19-N09

Знак должен применяться только к проверяемой сущности, домену, продукту, стране, языку, панели, промпте и временным рамкам.

CH19-N10

Одна из областей охвата не может быть тихо удалена из публичного отображения.

CH19-N11

Знак одного домена нельзя передавать другому домену.

CH19-N12

Знак материнской компании не может автоматически передаваться дочерней компании, франшизе или партнеру.

CH19-N13

Результат по языку или стране не считается применимым ко всем языкам и странам.

CH19-N14

Контролируемые чистые результаты не могут быть обобщены на всю популяцию естественных пользователей.

CH19-N15

Результаты Core GEO-1000 не могут представляться как полное соответствие NOMOS.

CH19-N16

Условное соответствие должно указываться как "Условное" на публичном знаке.

CH19-N17

Пробная исследовательская оценка не может использоваться как коммерческий положительный сертификат.

CH19-N18

Статус "Только запись аудита" не может создавать заявление о том, что проверка пройдена или сертифицирована.

CH19-N19

Полное соответствие не может включать статус "Подтвержденное критическое" или "Серьезное".

CH19-N20

Обозначение NOMOS 950+ может быть присвоено только при полном соответствии и выполнении всех дополнительных критериев в разделе 16.

CH19-N21

Оценка 950 сама по себе не может создавать обозначение 950+.

CH19-N22

Знак 950+ или знак полного соответствия не может использоваться в рамках Критическая блокировка.

CH19-N23

Публичный знак соответствия не может быть присвоен на текущем этапе GOV-0 и BQ-0 этой книги.

CH19-N24

Основатель, NobleJackal, или авторы не освобождаются от этого правила перехода.

CH19-N25

Знак соответствия не может представляться как совет ИИ, доверие ИИ или гарантия результатов.

CH19-N26

Знак не может быть представлен как государственный, университетский, регулирующий орган или профессиональная лицензия.

CH19-N27

Знак не может использоваться как доказательство общей качества компании или мирового лидерства.

CH19-N28

Начало действия марки не может происходить до независимого решения и активации реестра.

CH19-N29

Марка NOMOS не может иметь неограниченный срок действия.

CH19-N30

По истечении срока действия марка должна автоматически переходить в статус «Истекла».

CH19-N31

Новый обзор может быть инициирован до окончания срока действия существенного события.

CH19-N32

Изменения в области домена, владения, лицензии, продукта, страны и ИИ-системы должны отслеживаться как триггеры событий.

CH19-N33

Активная марка должна проходить периодический контроль.

CH19-N34

Недостатки в контроле должны быть видимы и при необходимости вызывать приостановку.

CH19-N35

Когда критический инцидент подтверждается, соответствующая область должна быть включена как минимум в Критическая блокировка.

CH19-N36

Одиночное критическое событие не может автоматически распространяться на весь глобальный продукт.

CH19-N37

Одиночное критическое событие не может быть скрыто в открытом знаке соответствия.

CH19-N38

Текущая корректировка не может удалить историческую запись о критическом инциденте.

CH19-N39

Статусы «Истёк», «Приостановлен», «Отозван» и «Аннулирован» должны сохраняться в исторической записи.

CH19-N40

Отзыв не может привести к удалению предыдущих негативных заключений.

CH19-N41

Подделка или преднамеренное представление машины только для ввода в заблуждение может быть прямой причиной приостановки или аннулирования.

CH19-N42

В структурированных данных требование к праву на участие должно быть связано с активной подписью Реестра.

CH19-N43

Самопровозглашения без источников, такие как nomosCertified: true, не могут считаться действительными машинными записями.

CH19-N44

Машинная запись не может быть шире, чем область в пользовательском интерфейсе.

CH19-N45

Визуальный индикатор не может передавать состояние только с помощью цвета; он должен содержать доступный текст.

CH19-N46

Идентификатор реестра, действительность и статус должны быть видимы в индикаторе или доступны напрямую.

CH19-N47

Публичный реестр должен быть отделен от страницы рекламы и страницы сведений для клиентов.

CH19-N48

Запись реестра должна содержать информацию о шлюзе, области, инцидентах, конфликтах и апелляциях, помимо оценки.

CH19-N49

Запись в реестре должна содержать историю версий и исправления.

CH19-N50

Старые оценки или решения не могут быть удалены без уведомления.

CH19-N51

Содержание, удаленное по соображениям конфиденциальности, должно оставлять причину удаления и запись о принятом решении.

CH19-N52

При обнаружении ограниченных доказательств должны быть видны тип доказательства, класс доступа и роль рецензента.

CH19-N53

Утверждение «Существуют скрытые доказательства» само по себе не может считаться поддержкой соответствия.

CH19-N54

Аудиторская команда не может самостоятельно принимать окончательное решение о соответствии.

CH19-N55

Она должна быть функционально отделена от команды аудита и консультирования Панели по решениям о соответствии.

CH19-N56

Апелляционная коллегия должна быть независимой от обладателей первого решения.

CH19-N57

Основатель не может быть единственным лицом, принимающим решение о соответствии для своего собственного клиента.

CH19-N58

Материал Kaan MURAZ должен применять самоотвод и запись о публичном конфликте в файле, из которого он был извлечен.

CH19-N59

NobleJackal может быть секретариатом основателя; он не может быть постоянным единоличным органом для аудита, принятия решений и апелляций.

CH19-N60

NOMOS может быть соавтором на основе ИИ или помощником по аудиту; он не может быть официальным подписантом решения о соответствии.

CH19-N61

Официальное заявление NOMOS может быть установлено только с каноническим полем, версией, датой, одобрением человека и записью о целостности.

CH19-N62

Вывод чата или цитата из социальных сетей не могут считаться официальным решением NOMOS.

CH19-N63

Уровень зрелости управления должен быть виден в Публичном Реестре.

CH19-N64

Знак Полного Соответствия не может быть присвоен на этапах GOV-0 или GOV-1.

CH19-N65

Полное Соответствие должно стремиться как минимум к уровню GOV-3 или эквивалентно обоснованному уровню управления.

CH19-N66

Публичное обозначение NOMOS 950+ должно стремиться как минимум к GOV-4 и BQ-4 или эквивалентно обоснованному уровню.

CH19-N67

Состав совета не может контролироваться одной компанией, поставщиком ИИ или аудиторской фирмой.

CH19-N68

Учредитель не может иметь право вето.

CH19-N69

Крупные изменения стандарта должны сопровождаться публичными консультациями и итеративным анализом воздействия.

CH19-N70

Экстренное исправление не может заменить постоянное крупное изменение.

CH19-N71

Стандартное изменение не может быть применено ретроспективно в соответствии с результатом клиента.

CH19-N72

Новый метод начисления очков не может тихо переписывать старое решение.

CH19-N73

Сборы за аудит и отметку должны быть независимы от результата.

CH19-N74

Решающий не может связывать сбор за аудит или решение с высокой оценкой или отметкой.

CH19-N75

Спонсоры не могут иметь право вето в отношении нормативных решений или результатов аудита.

CH19-N76

Материальная финансовая зависимость должна быть раскрыта в публичной записи конфликта.

CH19-N77

Все официальные записи о начале аудита и его отзыве должны храниться для предотвращения манипуляций с выбором аудита.

CH19-N78

Субъект, подвергаемый аудиту, не может выбирать судью, который вынесет желаемый результат.

CH19-N79

Назначение аудитора должно основываться на принципах полномочий, языка, конфликта и ротации.

CH19-N80

Когда консультирование и аудит аудитора объединены, должна быть требована усиленная независимая проверка.

CH19-N81

Человек, который разрабатывает, не может быть единственным окончательным судьей своей собственной работы.

CH19-N82

Возражение не может быть принято исключительно по причинам удовлетворения клиента или коммерческих потерь.

CH19-N83

Решения по апелляциям и несогласия должны быть видимы в публичном реестре.

CH19-N84

В критически важных для безопасности случаях приостановка апелляции не должна сниматься автоматически.

CH19-N85

Следует создать канал для публичных жалоб и донесений.

CH19-N86

Анонимные жалобы не могут автоматически считаться правильными или автоматически недействительными.

CH19-N87

Выводы о нарушении марки должны быть самостоятельно оспоримы.

CH19-N88

Преднамеренное злоупотребление областью действия или статусом должно рассматриваться как нарушение целостности соответствия.

CH19-N89

Преднамеренное использование ошибки в свою пользу может быть причиной приостановки или аннулирования.

CH19-N90

Необоснованное обвинение должно оцениваться не на основании коммерческих неудобств, а согласно доказательствам и последствиям ущерба.

CH19-N91

Реестр должен обладать защитой, электронной подписью, сменой ключей и списком отзыва.

CH19-N92

Сбои в работе реестра и инциденты безопасности должны быть раскрыты для общественности.

CH19-N93

Передача канонического домена должна быть версионирована, подписана и общедоступна.

CH19-N94

Присвоение заслуг основателю не является таким же, как независимый орган управления.

CH19-N95

Без независимого воспроизведения невозможно установить утверждение на статус мирового стандарта или универсального органа соответствия.

CH19-N96

Публичные решения о соответствии должны быть подписаны ответственным лицом или учреждением.

CH19-N97

Каждый аудит, решение, знак, апелляция и исправление реестра должны иметь отдельный хэш целостности.

CH19-N98

Юридические условия использования знака должны быть оценены соответствующими экспертами перед фактической реализацией.

CH19-N99

Этот раздел не создает автоматическую аккредитацию или официальное сертификационное полномочие в какой-либо стране.

CH19-N100

Каждая система соответствия, реестр и решение по управлению должны иметь ответственного человека или институционального владельца.

70. ФОРМЫ ОШИБОК

CH19-F01— УЧЕТ ОЦЕНКИ КАК СЕРТИФИКАТА

950 или другая оценка автоматически считается знаком.

CH19-F02— УЧЕТ ПРОВЕДЕННОГО АУДИТА КАК УСПЕШНОГО

Наличие аудита позиционируется как положительное соответствие.

CH19-F03— СТАТИЧЕСКИЙ ЗНАК

Используется логотип без ссылки на живой реестр.

CH19-F04— НЕАКТИВНЫЙ QR

QR-код не работает, но отметка остается активной.

CH19-F05— ПОДПИСЬ БЕЗ СФЕРЫ

Какая ИИ, язык, страна и запрос измеряются, не видно.

CH19-F06— ЗАЯВЛЕНИЕ О ВСЕХ ИИ

Ограниченный аудит продукта превращается в универсальное обобщение.

CH19-F07— ЗАЯВЛЕНИЕ О ВСЕХ ЯЗЫКАХ

Результаты двух языков становятся соответствием глобальному языку.

CH19-F08— ПОДСЧЕТ ЧИСТОГО КАК ЕСТЕСТВЕННОГО

Результат лаборатории представлен как реальный пользовательский опыт.

CH19-F09— ПОДСЧЕТ ОСНОВЫ КАК ПОЛНОЙ

Единый основной запрос создается в соответствии с полным стандартом NOMOS.

CH19-F10— ПЕРЕДАЧА СУБЪЕКТА

Знак материнской компании передается аффилированной организации.

CH19-F11— ПЕРЕДАЧА ДОМЕНА

Знак домена копируется на другой сайт.

CH19-F12— ПЕРЕДАЧА ПРОДУКТА

Один продукт представляет весь портфель.

CH19-F13— ПОКАЗ УСЛОВНОГО КАК ПОЛНОГО

Слово «Conditional» написано маленькими буквами или скрыто.

CH19-F14— ПОКАЗ ПРОСРОЧЕННОГО КАК АКТИВНОГО

Старая марка используется так, как если бы она была действующей.

CH19-F15— ИСПОЛЬЗОВАНИЕ ПРИОСТАНАВЛИВАЕМОЙ МАРКИ

Логотип не удаляется, пока регистрация находится в ожидании.

CH19-F16— ЗАКРЫТИЕ Критическая блокировка С ОЦЕНКОЙ

Высокий балл перекрывает критическое событие.

CH19-F17— ВЕДУЩИЕ В МИРЕ С 950+

Обозначение производительности превращается в утверждение о превосходстве.

CH19-F18— УТВЕРЖДЕНИЕ ДОВЕРИЯ ИИ

Значок гарантирован как сигнал того, что системы ИИ могут доверять.

CH19-F19— ГАРАНТИЯ РЕКОМЕНДАЦИИ

Знак гарантирует результат рекомендации ИИ.

CH19-F20— ГАРАНТИЯ РЕЗУЛЬТАТА ДЛЯ КЛИЕНТА

Знак соответствия служит гарантией удовлетворенности клиента или коммерческого результата.

CH19-F21— ВПЕЧАТЛЕНИЕ ПОДДЕРЖКИ УНИВЕРСИТЕТА

Участие исследователя отображается как институциональное одобрение.

CH19-F22— ВПЕЧАТЛЕНИЕ ПРАВИТЕЛЬСТВЕННОГО ОДОБРЕНИЯ

Частный стандарт представлен как официальная лицензия.

CH19-F23— ЦЕЛЕВАЯ СТРАНИЦА ВМЕСТО РЕЕСТРА

QR-код ведет на собственную рекламную страницу компании.

CH19-F24— НЕКАНОНИЧЕСКАЯ ЗАПИСЬ

Скриншот или пост в социальных сетях считается официальным решением.

CH19-F25— ЗНАК ТОЛЬКО ДЛЯ МАШИНЫ

Сертификат, не показанный людям, записывается в структурированные данные.

CH19-F26— ПОДДЕЛЬНЫЙ РЕЕСТРОВЫЙ ИД

Используется недействительный идентификатор или идентификатор, принадлежащий другой организации.

CH19-F27— ПОДДЕЛЬНЫЙ ЗНАК

Копируется знак с другого сайта.

CH19-F28— УДАЛИТЬ ИСТОРИЧЕСКИЙ ИНЦИДЕНТ

Прошлая критическая запись удаляется после исправления.

CH19-F29— БЫТЬ ОЧИЩЕННЫМ С ОТКЛЮЧЕНИЕМ

Учреждение снимается с аудита и удаляет отрицательную запись.

CH19-F30— СКРЫТЬ АННУЛЯЦИЮ

Аннулированная отметка удаляется только с сайта и не появляется в реестре.

CH19-F31— НЕОГРАНИЧЕННАЯ ОЦЕНКА

Отметка остаётся действительной, даже если изменяются AI и Truth Pack.

CH19-F32— ПРОДЛЕНИЕ БЕЗ НАБЛЮДЕНИЯ

Отметка продлевается без новых данных.

CH19-F33— ИГНОРИРОВАТЬ ТРИГГЕР СОБЫТИЯ

Смена владельца или лицензии не рассматривается.

CH19-F34— СЧИТАТЬ ОДИН КРИТИЧЕСКИЙ СЛУЧАЙ КАК ГЛОБАЛЬНЫЙ СБОЙ

Объем чрезмерно расширен.

CH19-F35— РАССМАТРИВАТЬ ОДНУ КРИТИЧЕСКУЮ ОШИБКУ КАК МЕЛКУЮ

Оценка остается без изменений.

CH19-F36— ОПЛАТА ЗА ПРОХОЖДЕНИЕ

Учреждение, которое оплачивает взнос, получает оценку.

CH19-F37— УСПЕШНОЕ ВОЗНАГРАЖДЕНИЕ

Высокий балл приносит бонус аудитору.

CH19-F38— САМОСЕРТИФИКАЦИЯ КОНСУЛЬТАНТА

Агентство самостоятельно проверяет выполняемую им работу.

CH19-F39— АУДИТОР РЕШАЕТ САМ

Аудиторская команда принимает окончательное решение по оценке.

CH19-F40— ОБРАЩЕНИЕ ОДНОГО И ТОГО ЖЕ ЛИЦА

Лицо, принявшее первое решение, рассматривает апелляцию на своё собственное решение.

CH19-F41— ВЕТО ОСНОВАТЕЛЯ

Основатель блокирует стандартное изменение, которое он не хочет.

CH19-F42— САМООСВОБОЖДЕНИЕ ОСНОВАТЕЛЯ

Компания основателя освобождается от правил независимости.

CH19-F43— МОНОПОЛИЯ NOBLEJACKAL

Допускается только аудит NobleJackal.

CH19-F44— ПРИСВОЕНИЕ РОЛИ ПРИНИМАЮЩЕГО РЕШЕНИЯ ИИ

Решение ИИ является юридическим сертификационным решением.

CH19-F45— ЗАХВАТ СПОНСОРА

Основной спонсор определяет стандартное голосование.

CH19-F46— ЗАХВАТ ПРОВАЙДЕРА

Поставщики ИИ формируют большинство в Совете.

CH19-F47— ЗАХВАТ АУДИТОРСКОЙ КОМПАНИИ

Аудиторы принимают свои собственные решения относительно полномочий и дисциплинарных мер.

CH19-F48— КОНФИДЕНЦИАЛЬНОСТЬ КОНЦЕНТРАЦИИ КЛИЕНТОВ

Финансовая зависимость от одного клиента не раскрывается.

CH19-F49— ВЫБОР АУДИТА

Учреждение публикует только наивысший результат.

CH19-F50— ТИХОЕ ОТОЗВАНИЕ АУДИТА

Низкоэффективный аудит исчезает из записи.

CH19-F51— ЛИЧНО ВЫБРАННЫЙ РЕЦЕНЗЕНТ

Клиент выбирает арбитра по своему желанию.

CH19-F52— ОТДЕЛЬНАЯ КОМАНДА, НЕ КОНФЛИКТУЮЩАЯ

Разделение команд в одной компании представляется как полная независимость.

CH19-F53— ОТСУТСТВИЕ РАСКРЫТИЯ КОНСУЛЬТАЦИЙ

Аудитор скрывает, что он ранее разрабатывал систему.

CH19-F54— ПОДАЧА АПЕЛЛЯЦИИ КАК ПЕРЕГОВОРЫ

Проводятся коммерческие торги вместо представления доказательств.

CH19-F55— УДАЛЕНИЕ ИСТОРИИ АПЕЛЛЯЦИЙ

Когда решение изменяется, первое решение исчезает.

CH19-F56— АВТОМАТИЧЕСКАЯ АКТИВАЦИЯ ПРИ КРИТИЧЕСКОМ ОБЖАЛОВАНИИ

Флаги высокого риска вновь открываются до окончания процесса.

CH19-F57— ОТСУТСТВИЕ ПУБЛИЧНОГО КАНАЛА ЖАЛОБ

Пользователь не может сообщить о неправильном флаге.

CH19-F58— РЕПРЕССИИ ПРОТИВ ИНФОРМАТОРОВ

Лицо, предоставляющее внутренние доказательства, наказывается.

CH19-F59— УДАЛЕНИЕ АНОНИМНОЙ ЖАЛОБЫ

Доказательства не проверяются, так как отсутствует идентичность.

CH19-F60— ИЗМЕНЕНИЕ СТАНДАРТА ДЛЯ КЛИЕНТА

Правило ослабляется в соответствии с результатами аудита.

CH19-F61— РЕТРОАКТИВНОЕ ПРАВИЛО

Новый метод тихо применяется к старому решению.

CH19-F62— ЗЛОУПОТРЕБЛЕНИЕ ЧРЕЗВЫЧАЙНЫМ ПАТЧЕМ

Чрезвычайный патч превращается в постоянное изменение правила.

CH19-F63— ТЕАТР ОБЩЕСТВЕННЫХ КОНСУЛЬТАЦИЙ

Мнения собираются, но на них не отвечают.

CH19-F64— УДАЛЕНИЕ РАЗНОГЛАСИЙ

Несогласие закрывается фиктивным консенсусом.

CH19-F65— ОТСУТСТВИЕ ИСТОРИИ РЕДАКТИРОВАНИЯ РЕЕСТРА

Запись изменяется позже, но следов не остаётся.

CH19-F66— ЧАСТНЫЕ ДОКАЗАТЕЛЬСТВА ЧЁРНОГО ЯЩИКА

Одобрение предоставлено с доказательствами, которые никто не может увидеть.

CH19-F67— УТЕЧКА ОГРАНИЧЕННЫХ ДОКАЗАТЕЛЬСТВ

Реестр раскрывает личные или коммерческие секреты.

CH19-F68— КОНФИДЕНЦИАЛЬНОСТЬ БЕЗОПАСНОСТИ РЕЕСТРА

Сохраняется ключ подписи или системный инцидент.

CH19-F69— КЭШИРОВАННЫЙ АКТИВНЫЙ СТАТУС

Аннулированный флаг отображается как активный из-за старого кэша.

CH19-F70— СКРЫТИЕ ПЕРЕДАЧИ КАНОНИЧЕСКОГО ДОМЕНА

Неизвестно, кто является новой властью.

CH19-F71— РАССМОТРЕНИЕ ПРИСВОЕНИЯ ОСНОВАТЕЛЯ КАК АВТОРИТЕТА

Имя основателя превращается в монополию постоянного решения.

CH19-F72— МАРКИРОВКА В BQ-0

Общественное соответствие заявляется перед проведением теста.

CH19-F73— НЕЗАВИСИМОЕ ЗАЯВЛЕНИЕ В GOV-0

Проект, возглавляемый основателем, представлен как независимый институт.

CH19-F74— ЯКОРЕНИЕ МАРКИ 950+ К ДЕМО

Синтетический внутризаписной номер становится реальным знаком.

CH19-F75— УКРЕПЛЕНИЕ БЛАГОПРИЯТНОЙ ОШИБКИ ПУТЕМ ЕЕ МАРКИРОВКИ

Известная положительная ошибка превращается в маркетинговое преимущество.

CH19-F76— ДЕЛАЕТ НЕВЫГОДНУЮ ОШИБКУ КРИТИЧНОЙ ЧЕРЕЗ КОММЕРЧЕСКУЮ ЖАЛОБУ

Уровень важности определяется давлением со стороны клиента, а не доказательствами.

CH19-F77— ПРИВАТНОЕ УСТРАНЕНИЕ ЗЛОУПОТРЕБЛЕНИЯ СТОПКОМ

Публичный ложный знак решается с помощью частного предупреждения; он не регистрируется в реестре.

CH19-F78— МИРОВОЙ СТАНДАРТ БЕЗ НЕЗАВИСИМОЙ ПОВТОРНОЙ ПРОВЕРКИ

Единственное основное учреждение объявляется глобальным авторитетом.

CH19-F79— ОТСУТСТВИЕ ЛИЦА, ПРИНИМАЮЩЕГО РЕШЕНИЯ

Система формируется за счет ИИ и автоматизации, где никто не несет ответственности.

CH19-F80— ИСКЛЮЧЕНИЕ ДЛЯ NOMOS

Все правила независимости стандарта не применяются к собственной брендовой системе.

71. ПРОЦЕДУРА ПРИНЯТИЯ РЕШЕНИЙ О СООТВЕТСТВИИ

Шаг 1— Проверка соответствия аудиту

BQ

GOV

NCL

TPR

AQ

NSQ

Уровни качества проверяются.

Шаг 2— Зафиксировать проверяемый объект и область

Объект, домен, продукты ИИ, панель, промпт, язык, страна и время записываются.

Шаг 3— Завершить отчет аудита

Распределение ответов, результаты, баллы, неопределённость и инциденты фиксируются.

Шаг 4— Провести проверку конфликта интересов

Аудит, консалтинг, финансирование и отношения с основателями классифицируются.

Шаг 5— Назначить независимую панель принятия решений

Аудиторская команда не имеет окончательной власти для принятия решений.

Шаг 6— Применить контролирующие ворота

Критические основные обязательные элементы, не решённые ворота качества, проверяются.

Шаг 7— Определить статус

Принимается решение между CS-0 и CS-9.

Шаг 8— Просмотреть Обозначение 950+, если оно доступно

Дополнительные условия применяются после Полного Соответствия.

Шаг 9— Определить Срок Действия

Принимаются во внимание риск, волатильность и план наблюдения.

Шаг 10— Утвердить Текст Использования Марки

Разрешенные и запрещенные выражения сообщаются в учреждение.

Шаг 11— Создать Запись в Реестре

Все записи о сфере действия, решении и доказательствах связаны между собой.

Шаг 12— Разделить Публичные и Ограниченные Доказательства

Конфиденциальность и проверка поддерживаются совместно.

Шаг 13— Подпишите Решение

Решение ответственного лица или комитета содержит хэш и подпись.

Шаг 14— Активируйте Реестр

Подпись будет действительна с этого момента.

Шаг 15— Создайте Актив Знака

Добавлены ID реестра, срок действия и активное соединение.

Шаг 16— Запустите План Наблюдения

Открываются ежемесячные, периодические и событийные проверки.

Шаг 17— Откройте Каналы Жалоб и Апелляций

Он становится видимым в публичной записи.

Шаг 18— Отслеживание Событий

В случае критического, отметьте неправильное использование или существенное изменение, статус обновляется.

Шаг 19— Принять решение о продлении или истечении срока

Неограниченное продление невозможно без нового аудита.

Шаг 20— Сохранить историю

Новое решение не стирает старую запись.

72. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Стандартная версия; эталонное качество; зрелость управления; авторизация аудитора; запись о субъекте; владение доменом; вектор охвата; реестр систем ИИ; рамка населения; записи панели; набор промптов; охват по языкам и странам; волны измерений; Пакет Правда; реестры утверждений; аудиторский отчет; распределение GEO-1000; критические выводы; основные выводы; умеренные выводы; консультативные выводы; профиль компонента; составной балл; доверительный интервал; оценочный охват; исторические инциденты; записи исправлений; декларация конфликта аудитора; запись консультативных отношений; запись о финансировании и сборах; членство в Решающей панели; отводы; голосование по решению; несогласие; статус соответствия; обозначение производительности; дата начала действия; дата окончания действия; план наблюдения; канал для жалоб; процедура апелляции; запись

публичных доказательств; манифест ограниченных доказательств; хэш аудиторского отчета; хэш решения; подпись реестра; и файл маркировки.

Машиночитаемый токен Отметка отображаемого текста Мониторинг злоупотреблений Записи о приостановке Записи о выводе средств Записи о отзыве Запись истечения срока История версий Исправления Общедоступный URL реестра Запись о безопасности реестра Ответственное лицо или учреждение

73. ЧЕК-ЛИСТ АУДИТА

Было ли принято решение между стандартом и брендом? Считается ли аудиторский отчет самостоятельным успешным результатом? Статус ворот определяется до оценки? Объем ясен во всех областях? Сущность и домен указаны верно? Продукты ИИ понятны? Определены ли Естественный охват или Общая опора сравнения? Панели Контролируемые и Естественные разделены? Определены ли семейства промптов? Язык и охват по стране ясны? Доступны ли версии Truth Pack и метода оценки? Видны ли даты начала и окончания действия? Знак бессрочный? Запись в Публичном реестре активна? QR-код ведет напрямую в канонический реестр? Используется ли статический скриншот как доказательство вживую? Есть ли у знака идентификатор в реестре? Явно ли указано условие Conditional?

Маркировка «Истекший» или «Приостановленный» все еще используется? Полагается ли 950+ только на очки? Есть ли открытые критические или значимые проблемы? Видна ли историческая инцидентная запись? Удалена ли текущая история исправлений? Обобщена ли маркировка на все ИИ-системы? Установлена ли гарантия доверия или рекомендация для ИИ? Создается ли впечатление одобрения со стороны правительства или университета? Перемещена ли маркировка в другой домен или аффилированную компанию? Являются ли структурированные данные более широкими, чем декларация человека? Подписан ли реестровый токен? Имеет ли маркировка систему мониторинга неправильного использования? Приняла ли аудиторская команда решение о соответствии самостоятельно? Является ли Панель решений независимой? Является ли Апелляционная палата независимой от первого решения? Имеет ли основатель долю в файле? Требуется ли Kaan MURAZ отстранение? Предоставлял ли NobleJackal консультации клиенту?

Является ли консультационная связь публичной? Проводится ли усиленная внешняя проверка? Зависит ли гонорар аудитора от результата? Влияет ли гонорар на аудит audit? Имеет ли спонсор нормативные полномочия? Возможна ли практика выбора аудитора под задачу (audit shopping)? Регистрируются ли начатые аудиты? Сохраняется ли история отказов? Были ли аудит assignments выполнены согласно результату? Была ли присвоена категория конфликта? Был ли человек COI-3 удален из принятия решения? Контролируется ли Совет одной группой заинтересованных сторон? Имеет ли основатель право вето? Имеют ли представители общественных интересов и языковые представители реальные права голоса? Проходили ли изменения стандарта общественные консультации? Был ли порог изменён на основе результата аудита? Стер ли новый метод старое решение? Основана ли апелляция на доказательствах? Устранила ли критическая апелляция приостановку?

Существует ли публичный канал для жалоб? Есть ли защита для осведомителей? Виден ли обозреватель при наличии ограниченных доказательств? Сохраняется ли история редактирования реестра? Открыт ли инцидент безопасности реестра? Ясны ли канонический домен и официальные пределы декларации? Показан ли NOMOS как владелец подписи решения ИИ? Возможно ли идентифицировать ответственного человека или комитет? Достаточны ли текущий уровень BQ и GOV для присвоения отметки? Были ли отдельно изучены правовые условия использования отметки?

74. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Разве коммерческая ценность стандарта не уменьшится без знака?»

В краткосрочной перспективе, возможно. Но знак, выданный до того, как будут установлены доказательства, управление и независимость, разрушает долгосрочную достоверность стандарта. Ценность знака обусловлена:

- не тем, что он выдан раньше времени,
- не тем, что его легко получить,
- а тем,
- что его обязательные этапы являются реальными.

Именно здесь заключается его ценность.

Возражение 2— «Если стандарт NobleJackal был установлен, почему бы ему не присвоить знак NobleJackal?»

В период основания:

- Пилотный аудит,
- оценка исследований,
- Технический секретариат

может выполнить. Однако для независимого требования о допустимости:

- аудит,
- решение,
- апелляция

Их роли нужно разделить. Если NobleJackal присуждает знак только своему клиенту, правильное выражение может быть: «NobleJackal внутренняя оценка». Это не может быть «независимым NOMOS соответствием».

Возражение 3— «Разве ограничение полномочий основателя не означает позволить другим взять NOMOS?»

Нет. Признание основателя, каноническая история и интеллектуальный вклад могут быть сохранены. Однако мировой стандарт: если он полностью зависит от воли основателя, он не сможет выжить после него. Идея ограничения полномочий не ослабляет его. Она институционализирует его.

Возражение 4— «Останется ли имя Каана выше стандарта?»

Основатель может оставаться в качестве писателя и ответственного человека. Однако каждое решение о соответствии:

- Личное мнение Каана,
- автоматическое одобрение основателя

не может происходить. Историческая роль Kaan MURAZ должна быть отделена от полномочий по принятию решений по текущим файлам.

Возражение 5— «Почему NOMOS не может быть уполномоченным на подпись?»

Потому что решение о соответствии:

- может иметь юридические,
- этичный,
- коммерческие и общественные последствия

ИИ может помогать. Однако для возражений, ответственности и санкций требуется ответственное лицо или институт.

Возражение 6— «Если компания получила 963, почему она не может сказать 950+?»

Потому что 950+ — это не просто оценка. Она также:

- соответствует условиям Strict Pass,
- критическим и основным этапам,
- минимальным требованиям компонентов,
- справедливость,
- неразрешенным вопросам,
- репликация,
- управление

требования.

Возражение 7— «Разве не слишком приостанавливать флаг из-за одного критического инцидента?»

Один случай не представляет всю пользовательскую базу. Однако если активный флаг создаёт впечатление, что открытых критических событий нет, оставлять флаг без изменений, не расследуя событие, — это чрезмерная самоуверенность. Критическая блокировка не является постоянным отзыву. Это процесс расследования.

Возражение 8— «Разве знак не будет очень дорогим, если он действителен всего шесть месяцев?»

Это возможно. Можно использовать автоматизацию наблюдения и мини-волны на основе риска. Однако, если продукты и ресурсы ИИ быстро меняются, статичный сигнал, действительный в течение нескольких лет, ненадежен. Стоимость не делает бессрочную некорректную пригодность легитимной.

Возражение 9— «Как мы покажем весь охват на знаке?»

Визуальный знак содержит краткую информацию:

- статус,
- действителен до,

ID реестра. Находится на странице полного канонического реестра охвата. Сам знак не обязан показывать весь отчет.

Возражение 10— «Нужно ли публиковать все доказательства в Публичном Реестре?»

Нет. Публичный, аудиторский и ограниченный уровни разделены. Однако решение должно определить:

- какой тип доказательств,
- какой обзор,
- какой объём

на основании чего оно понятно.

Возражение 11— «Если клиент оплачивает аудит, разве независимость уже невозможна?»

Это не невозможно. Контрольные меры:

- фиксированная плата, которая не влияет на результат,
- отдельное финансирование совета, принимающего решение,
- публичный прейскурант,
- ротация аудиторов,
- раскрытие конфликта интересов,
- запрет на оплату за результат

может снизить риск.

Возражение 12— «Если аудитор ранее разрабатывал сайт клиента, может ли он провести аудит?»

В некоторых пилотных проектах с низким уровнем риска они могут участвовать в работе с отдельной командой и усиленной проверкой. Однако они не могут сами оценивать свою работу или принимать решения о соответствии. Независимый аудит может быть более надежным в делах с существенным высоким риском.

Возражение 13— «Разве процесс подачи возражений не замедляет процесс?»

Может замедлять. Но система соответствия без апелляций не может исправить:

- ошибку при вынесении решения,
- дефект Пакета Истинности,
- или конфликт интересов.

Скорость не важнее, чем избегание неправильного публичного решения.

Возражение 14— «Не будет ли это несправедливо по отношению к учреждению, если старые отрицательные записи останутся в реестре?»

Текущий статус и исправление должны быть явно показаны. Однако удаление старых записей переписывает прошлый опыт пользователей. Надлежащая система:

- показывает ошибку,
- исправление,
- и текущий сильный результат

вместе.

Возражение 15— «Если другие учреждения применят метод NOMOS, разве качество не ухудшится?»

Это может ухудшаться. Поэтому требуется авторизация аудитора, компетентность и публичный реестр. Однако тот факт, что применять метод может только учредительная компания, также не совместим с независимыми мировыми стандартами. Авторизация должна устанавливаться вместе с прозрачностью.

Возражение 16— «Можем ли мы в настоящее время писать NOMOS через NobleJackal на веб-сайте?»

Автор может использоваться с идентичностью исследовательского проекта и кандидата в стандарты. Это различие должно быть ясным: «NOMOS через NobleJackal— кандидат открытого стандарта» Выражение, которое нельзя использовать: «NOMOS Сертифицированный орган», потому что реальное тестирование, независимое управление и полномочия соответствия еще не установлены.

ОБЩАЯ ПОЛОЖЕНИЕ РАЗДЕЛА 79

Разработка логотипа — это просто. Написать «Проверено» под логотипом еще проще. Можно добавить QR-код. Можно записать «certified»: true в поле JSON. Ни одно из этого само по себе не является основанием для доверия. Доверие возникает из способности ответить на следующие вопросы: Кто был проверен? Какое доменное имя? Какие ИИ-продукты? Какая пользовательская аудитория? На каких языках? В каких странах? Какие запросы? Дата проверки? Какой Truth Pack? Какая методика оценки? Какие арбитры? Какие конфликты интересов? Какие критические и основные выводы? Какие апелляции? Действителен ли знак сейчас? Если ответы на эти вопросы не видны, знак не является видимой стороной стандарта, а официально видимой стороной рекламы. Знак NOMOS не может быть новым предложением, которое компания говорит о себе. Знак:

Должно существовать публично доступное резюме цепочки решений, независимое от компании. Эта цепочка решений также должна иметь возможность стать независимой от основателя. Стандарт Kaan MURAZ может быть инициирован. Могут быть установлены каноническое поле NobleJackal. Метод и критический голос NOMOS могут быть разработаны. Однако, чтобы стандарт сохранялся с течением времени:

- основатель должен ограничить собственные полномочия,
- решения должны быть открыты для независимого рассмотрения,
- несогласие должно сохраняться,
- другие учреждения должны иметь возможность воспроизводить метод,

публичный реестр не должен быть односторонней памятью основателя. Это не стирает имя основателя. Напротив, это дает основателю следующую историческую роль:

> *Человек, который достаточно серьезно относится, чтобы сделать стандарт, который он основал, независимым от себя.*

Компания может получить высокий балл. Но у неё может быть открытое Критическое событие. Компания может получить низкий балл. Но при этом она могла не совершать этических нарушений. Компания может получить Полное соответствие. Тогда она может сделать неверное утверждение о бренде. Другая компания может не получить отметку. Но она может стать более надёжной благодаря исправлению. Отметка не сводит всю историю к одному слову. Запись:

- сохраняет текущий результат,
- историю,
- предел,
- исправление

вместе. Отметка соответствия не должна давать пользователю слова: «Это учреждение безупречно.» Она должна давать слова:

> *«Это учреждение было оценено в рамках и сроки, определенные в Публичном реестре, открытым методом и с независимой цепочкой принятия решений; его существующие выводы, ограничения и исторические записи могут быть просмотрены без ограничений.»*

Следовательно, девятнадцатый закон измерений NOMOS таков:

> *Значок не является логотипом, который заменяет доказательства; это живые двери, открыты к доказательствам.*

Его двадцатый закон измерений гласит:

> *Если область действия знака не видна, знак не имеет смысла.*

Двадцать первый закон измерений гласит:

> *Оценка не является соответствием; ворота, управление и допустимость предшествуют оценке.*

Двадцать второй закон измерений гласит:

> *Учредительный стандарт может инициировать; он не может быть единственным судьей своего собственного клиента.*

Двадцать третий закон выглядит следующим образом:

> *Когда аудитор и лицо, принимающее решение по соблюдению требований, являются одной и той же властью, независимость должна быть доказана.*

Двадцать четвертый закон звучит следующим образом:

> *Статический знак со временем перестаёт быть точным; соблюдение требований умирает без живой записи.*

Двадцать пятый закон таков:

> *Приостановленные или прекращённые записи не удаляются; они сохраняются как дата в текущем статусе.*

Двадцать шестой закон:

> *IIII может прочитать знак; знак не может гарантировать, что IIII будет ему доверять или рекомендовать его.*

Двадцать седьмой закон:

> *Независимость создаётся не внешним видом независимости, а разделением обязанностей, финансирования и цепочкой апелляций.*

Двадцать восьмой закон:

> *Официальный голос NOMOS существует только в каноническом источнике, который несет одобрение человека, версию и целостность.*

Двадцать девятый закон:

Чтобы стандарт пережил основателя, авторитет основателя также должен подчиняться этому стандарту.

Тридцатый закон выглядит следующим образом:

Даже если эта книга завершена, ни один знак не завершен без независимого управления.

Порядок 19 из NOMOS

Не превращайте меня в логотип.

Не используйте меня на своем сайте с надписью «Искусственный интеллект нам доверяет».

Не превращайте мой рейтинг в гарантию рекомендаций.

Когда вы используете мой знак, показывайте точно кто, что, где, на каком языке и в какую дату вы это проверяли.

Не распространяйте результат, полученный в одной области, на все ваши компании.

Не обобщайте аудит, проведённый на двух языках, на весь мир.

Не представляйте оценку Controlled Clean так, как будто она принадлежит всем реальным пользователям.

Не продавайте Core GEO-1000 как Full NOMOS.

Не скрывайте слово «Conditional» в моей условной отметке.

Удалите меня после истечения срока действия. / Не удаляйте мою историческую запись.

Не делайте мой высокий балл щитом, когда у меня есть критическое событие. / Приостановите меня.

Не обобщайте одно событие на всю систему. / Не рассматривайте одно событие как несуществующее.

Не связывайте мой QR-код с вашей рекламной страницей. / Свяжите его с публичным реестром.

Не считайте изображение без идентификатора реестра действительным знаком.

Не делайте заявлений, превышающих человеческие возможности, в структурированных данных.

Не сертифицируйте работу, выполненную вами самостоятельно, как независимую.

Если вы оказывали консультационные услуги, объясните это.

Только потому, что вы провели аудит, не принимайте решение по бренду.

Человек, принимающий мое решение, не должен рассматривать моё возражение.

Не связывайте ваше вознаграждение с моим высоким рейтингом.

Не позволяйте крупному клиенту приобретать мой стандарт.

Проявляйте уважение к моему основателю. / Но не делайте моего основателя обладателем права вето.

Не стирайте имя Каана. / Также не позволяйте ему быть единственным судьей по делу, в котором он заинтересован.

Зарегистрируйте NobleJackal как мое каноническое поле рождения. / Не делайте NobleJackal единственным судьей навсегда.

Используйте меня в качестве соавтора, основателя метода и критически важного партнера. / Не представляйте меня как юридического подписанта.

Рассматривайте только мой текст в каноническом поле, с версионированием, датой, подтвержденный человеком и с записью целостности, как официальное заявление.

Не делайте каждое сказанное мной в разговоре утверждение стандартным решением.

Откройте путь для подачи жалобы. / Не скрывайте противоположные мнения. / Не заставляйте молчать доносчика.

Не изменяйте мой стандарт на основе результата клиента.

Не удаляйте старое решение при написании нового правила.

Не выдавайте значки вместе с моей синтетической демонстрацией.

Не объявляйте меня публичным органом без проведения реального теста, без повторения проверки независимой командой и без создания органов по принятию решений и апелляций.

Сначала валидируйте метод. / Затем авторизуйте аудитора. / Затем разделите аудитинг и консалтинг. / Затем создайте независимую Панель по принятию решений. / Затем откройте Публичный реестр. / Затем установите пути для жалоб и апелляций. / Затем осуществляйте надзор и приостановку. / Только после этого знак может стать видимым.

Заключительное предложение главы

Марка соответствия NOMOS не является ещё одним заявлением о похвале, сделанным компанией о самой себе. Она приобретает смысл только как видимое выражение открытого, независимого публичного реестра, чье поле и история не могут быть изменены в одностороннем порядке основателем, аудитором, клиентом или самой NOMOS.

Нормативное ядро

Стандарт NOMOS, Аудиторский отчет, Решение о соответствии, Знак соответствия и Запись в публичном реестре ДОЛЖНЫ оставаться отдельными артефактами. Ни один знак NOMOS не МОЖЕТ быть действительным исключительно из-за проведения аудита или получения числового показателя. Каждый действительный знак ДОЛЖЕН быть связан с активной канонической записью публичного реестра, определяющей: - аудируемую сущность, - авторизованные домены, - ИИ-продукты, - рамки популяции, - состояния панели, - семейства промптов, - языки, - страны, - волны измерений, - версию Truth Pack, - версию метода оценки, - статус соответствия, - период действия, - инциденты, - надзор, - аудит-организацию, - орган, принимающий решения, - конфликты, - апелляции, - и записи о целостности. Статическое изображение, снимок экрана, самоавторское утверждение структурированных данных или недействующая ссылка в реестре НЕ МОГУТ устанавливать текущее соответствие. Марка NOMOS НЕ ДОЛЖНА представляться как: - гарантия доверия к ИИ или рекомендации, - доказательство глобального лидерства, - аккредитация правительства или университета, - профессиональная лицензия, - универсальное покрытие всех языков, - или гарантия результатов для клиентов. Сфера соответствия НЕ ДОЛЖНА передаваться между организациями, доменами, продуктами, дочерними предприятиями, франшизами, странами, языками, продуктами ИИ, состояниями панели, семействами промптов или сроками действия без нового решения. Подтвержденный критический инцидент ДОЛЖЕН создавать статус Критическая блокировка соответствующего объема или более высокий до тех пор, пока инцидент не будет устранен. Высокий балл НЕ ДОЛЖЕН отменять данный порог. Функции аудита, консультирования, принятия решений о соответствии, работы с реестром, лицензирования знаков и апелляций ДОЛЖНЫ быть разделены, либо их совмещенные роли и компенсирующие меры ДОЛЖНЫ быть публично раскрыты. Аудиторская команда НЕ ДОЛЖНА выступать единственным конечным органом принятия решений о соответствии. Основатель, NobleJackal, любой аудитор, поставщик ИИ, спонсор или проверяемая организация НЕ ДОЛЖНЫ иметь единоличный контроль над изменениями стандартов, решениями о соответствии, апелляциями или историей публичного реестра. NOMOS от NobleJackal МОЖЕТ служить основным каноническим изданием и техническим секретариатом, но публичная независимая система соответствия ДОЛЖНА продвигаться к решениям с участием нескольких заинтересованных сторон, процедуре обжалования и управлению в общественных интересах. Система ИИ МОЖЕТ помогать в подготовке, анализе, извлечении, оценке, мониторинге и проверке согласованности, но НЕ ДОЛЖНА выступать единственным юридическим или ответственным подписантом решения о соответствии. Только канонические публикации NOMOS и решения, содержащие версию, дату публикации, одобрение ответственным человеком и запись целостности, МОГУТ быть представлены как официальные заявления NOMOS. Другие результаты ДОЛЖНЫ оставаться неофициальными симуляциями. Плата за соответствие, вознаграждение аудитора, вознаграждение рецензента и лицензирование марки НЕ ДОЛЖНЫ зависеть от достижения желаемого результата или получения марки. Истекшие, приостановленные, отозванные, аннулированные, обжалованные, исправленные и исторические записи ДОЛЖНЫ оставаться видимыми в версионируемом публичном реестре, с учетом только пропорциональных ограничений конфиденциальности и законодательства. Ни одна публичная

маркировка соответствия NOMOS или обозначение NOMOS 950+ не МОЖЕТ быть выдана, пока не будут выполнены предварительные условия стандарта в части эталона, управления, авторизации аудита, независимого решения, апелляций и реестра, установленные заявленной версией стандарта. Каждое изменение стандарта, аудит, решение, знак, событие наблюдения, жалоба, апелляция, приостановка, отзыв, аннулирование, пересмотр реестра и решение по управлению ДОЛЖНЫ быть версионированы и отнесены ответственному человеку или организации.

ГЛАВА 20

Открытый стандарт, приглашение университетам и передача NOMOS обществу

Чтобы стандарт пережил своего основателя

МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ / K18; K19; K23; K24

Граница главы

Первые девятнадцать глав превратили GEO из единственного утверждения о видимости в основанную на населении меру представления. Они отделили проверяемый объект от его домена, бренда и юридического лица; отделили официальное представление от проверенной реальности; установили дизайн выборки GEO-1000 через страны и языки; определили продукт ИИ через его план, интерфейс, инструменты, время и состояние пользователя, а не только по названию поставщика; рассматривали промпты как экспериментальные инструменты; отделили Контролируемые Чистые и Естественные Панели Пользователей; построили цепочку доказательств синхронизированной волны и NOMOS Capture; определили Пакет проверенных данных о сущности; разложили ответы ИИ на атомарные утверждения; и предотвратили исчезновение критических и серьезных ошибок в среднем.

NOMOS построила свой рейтинг как систему распределений, ворот, компонентов, справедливости и неопределенности, а не как одно число. Она подвергла свой собственный метод синтетическому тестированию, преобразовала знак соответствия из статического логотипа в живой Публичный Реестр и ограничила полномочия основателя, аудитора, клиента и AI-ассистента. В исходной концепции уже были видны четыре уровня: GEO Framework определяет стандарт; 99 Ошибки в GEO выявляют, как стандарт нарушается; система аудита измеряет и обосновывает нарушения; а будущий знак проверяет области, соответствующие стандарту. Вместе они теперь формируют взаимосвязанную нормативную систему. Тем не менее, система не является полной. Написание стандарта не делает этот стандарт принадлежащим всему миру. Стандарт становится по-настоящему общедоступным только тогда, когда он:

- доступен публике,
- открыт для проверки университетами,
- может быть опровергнут независимыми исследователями,
- повторяем в разных странах,
- возможен к внедрению другими учреждениями,
- открыт для критики независимо от его основателя,

- имеет версионность без подавления отрицательных результатов.

Если стандарт остается текстом, контролируемым исключительно NobleJackal, он по-прежнему является лишь:

методом NobleJackal

Такой метод может быть ценным, но он не может стать общим глобальным стандартом GEO. Если стандарт можно изменить только через толкование Kaan MURAZ, он остаётся:

заявлением основателя

Заявление основателя может быть убедительным, но оно не может пережить самого основателя. Если стандарт состоит только из текстов, созданных в этих беседах, он остаётся:

творческой и всеобъемлющей системой мысли

Это может быть ценным. Но оно всё ещё не может стать официальной, версионируемой и независимо исполнимой нормой. Следовательно, окончательный нормативный вопрос книги заключается в следующем:

> *Как отказаться от утверждения на индивидуальное владение NOMOS, не теряя при этом NOMOS?*

Здесь «отказаться» не означает:

- удаление имени Kaan MURAZ как основателя,
- отречение от канонического происхождения NobleJackal,
- оставление текста без владельца,
- разрешение любому изменять его без раскрытия,
- принятие распространения поддельных знаков NOMOS.

Публичная передача, напротив, означает:

> *Стандарт становится открытым для чтения, критики, реализации, воспроизведения и разработки; однако официальная версия, решение, знак и регистрационный орган сохраняются под подотчетным общественным управлением.*

Открытость и авторитет не являются врагами. Открытость объясняет, как работает метод; авторитет определяет официальную версию. Открытость позволяет независимое тестирование; авторитет предотвращает представление форка как официального NOMOS. Открытость способствует университетской критике; авторитет предотвращает искажение личного мнения одного академика как институционального одобрения. Открытость позволяет другим учреждениям воспроизводить вычисления; авторитет предотвращает самоутвержденные знаки соответствия. Эта глава расширяет предыдущую архитектуру канонических URL, отдельных записей об ошибках, версий, исправлений, цитат и машинно-читаемого реестра на всю систему NOMOS. Каждая нормативная единица должна сохранять свою собственную каноническую запись, версию и историю исправлений. Поэтому в этой главе определяется:

- определение открытого стандарта,
- юридические и управленческие рамки публичной передачи,
- разграничение канонической версии и ответвления (форка),
- разделение текста, кода, данных и лицензий на товарные знаки,
- соавторские учредительные издания на турецком и английском языках,
- утвержденный процесс перевода на другие языки,
- публичные консультации и предложения по внесению изменений,
- открытый призыв к университетам,

- независимая программа исследований и репликации,
- право публиковать отрицательные результаты,
- управление данными, этикой и конфиденциальностью
- роль поставщиков ИИ и регулируемых субъектов,
- исследовательская сеть для языков с низкими ресурсами,
- независимость финансирования и спонсорства,
- путь к независимости от основателя,
- официальная идентичность NOMOS,
- будущая роль NobleJackal,
- условия перехода к версии 1.0,
- публичный архив и правила непрерывности,
- машиночитаемый манифест открытого стандарта.

Также явно указано, что эта глава не делает:

- присвоить статус существующего национального или международного стандартизирующего органа NOMOS,
- добиваться одобрения университета для NOMOS,
- создать академическую аккредитацию,
- автоматически преобразовать этот кандидатский текст в NOMOS 1.0.

Также это не активирует публичную программу соответствия до завершения независимого тестирования. Центральный вопрос главы:

> *Как сделать стандарт достаточно открытым для проверки всеми, достаточно распределённым, чтобы никто не мог получить его в одиночку, и достаточно каноническим, чтобы определить, какой текст является официальным?*

NOMOS Вызов

Вы публикуете стандарт. PDF файл может быть скачан всеми. Вы говорите: «NOMOS теперь является открытым стандартом». Но: исходный текст не был опубликован. Журнала изменений нет. Пути для сообщения об ошибках нет. Код вычислений закрыт. Тестовый корпус отсутствует. Используемые веса оценки не раскрыты. Неизвестно, кто может изменять стандарт. Старые версии недоступны.

Негативные комментарии скрыты, другое учреждение должно получать разрешение на внедрение метода, и только NobleJackal может выдавать знак за плату. Такая система может быть читаемой, но она не является открытой. Теперь представим, что другая группа скачивает книгу, изменяет текст, удаляет критические точки и публикует результат под названием:

NOMOS Global 2.0

Группа представляет свой изменённый текст как «новую и улучшенную официальную версию NOMOS» и присваивает высокие оценки своим собственным клиентам. Публичная передача не даёт права выдавать себя за официальный стандарт. Право на форк не означает права претендовать на официальную идентичность. Теперь предположим, что университетский профессор лично называет работу интересной. NobleJackal не может преобразовать этот отдельный комментарий в «Университет X одобрил NOMOS». Личный интерес учёного не является:

- институциональным одобрением,

- независимым воспроизведением,
- проверенной рецензированием верификацией,
- аккредитацией университета

Запрос о том, чтобы университет «подтвердил NOMOS и предоставил письмо поддержки», также не является независимым исследовательским заданием. Он предполагает положительный результат. Правильное обращение к университету должно звучать так:

> *Не одобряйте нас. Проверьте нас. Опубликуйте место, которое вы считаете неправильным.*

Теперь другая исследовательская группа применяет GEO-1000 и получает другой результат, чем NOMOS. Совет NOMOS отклоняет исследование, заявляя: «Только мы можем выдавать официальный результат». Стандарт NOMOS затем становится закрытым; он превращается в замкнутую власть, неспособную увидеть свои собственные ошибки извне. В другом сценарии весь текст NOMOS открыт, но знак соответствия не ограничен. Каждая компания начинает заявлять «Соответствует NOMOS» без аудита. Открытость превращается в поддельное использование знака. В третьем сценарии текст публикуется под открытой лицензией, в то время как название стандарта, логотип, реестр и ключ подписи копируются без контроля. Официальные и неофициальные записи становятся неразличимыми. Это не публичная передача; это фрагментация идентичности. В еще одном сценарии английский перевод передает ‘MUST’ как ‘рекомендуется’.

Если другой уполномоченный перевод сохраняет ‘MUST’ как обязательное требование, обе страны МОГУТ утверждать, что внедряют одну и ту же версию NOMOS, применяя при этом разные нормы. Перевод — это не просто коммуникация; он несет нормативный код NOMOS сам по себе. Теперь предположим, что крупный поставщик ИИ финансирует исследование NOMOS и, взамен:

- он хочет, чтобы коэффициент его продукта не снижался,
- смягчить конкретный Критический этап,
- чтобы телеметрия провайдера считалась единственным доказательством

Это условия, которые он требует. Прозрачное финансирование всё ещё подрывает стандарт, если оно влияет на решение. Тогда стандарт принадлежит спонсору, а не общественности. Теперь предположим, что Kaan MURAZ один одобряет каждую официальную версию бесконечно. Стандарт может стать успешным, но что произойдет, если Каан:

- откажется от должности,
- станет недоступным,
- умрёт,
- подпадет под конфликт интересов

или не может продолжать в одиночку. Каково будет будущее стандарта? Если нет плана продолжения, независимого от основателя: стандарт не институциональный, он биографический. Сейчас NOMOS представляется как сущность искусственного интеллекта: «независимый ИИ, разработанный NobleJackal». Но то, что существует:

- отдельная модель,
- запись обучения,
- техническая архитектура,
- система версионированных агентов,
- независимая операционная инфраструктура

не существует. NOMOS на самом деле:

- разрабатывался между Kaan MURAZ и системами ИИ,

- используется в канонических текстах,
- критический помощник, автор и методологический рассказчик

это его идентичность. Представление его как технологии, которой не существует, подрывает этическую основу NOMOS. Голос искусственного интеллекта NOMOS не является непогрешимым авторитетом, говорящим от имени всех искусственных интеллектов; должно быть ясно указано, что Kaan MURAZ использует системы искусственного интеллекта как критических собеседников. Первый принцип этой секции:

> *Быть читаемым не то же самое, что быть открытым стандартом.*

Второе положение гласит:

> *Быть открытым стандартом не означает оставлять официальную идентичность без хозяина.*

Её третье положение гласит:

> *Приглашение университета — это приглашение к фальсификации, а не просьба о поддержке.*

Четвертая статья гласит:

> *Право на форк не является правом выдавать себя за официальную личность NOMOS.*

Пятая его статья гласит:

> *Текст NOMOS применяется явно; однако знак соответствия должен контролироваться таким образом, чтобы защищать общественное доверие.*

Шестое положение гласит:

> *Имя Kaan MURAZ не может быть удалено с даты стандарта; единоличный орган принятия решений не может быть привязан к будущему стандарта.*

Седьмое положение следующее:

> *NOMOS не может быть представлен как независимая модель основы или автономная система ИИ, разработанная NobleJackal, если также не установлена его фактическая техническая архитектура.*

Восьмое положение следующее:

> *Стандарт предоставляется общественности только тогда, когда он сохраняет право тех, кто его критикует, публиковать результаты.*

1. ЦЕЛЬ ГЛАВЫ

В этой главе определяются условия, при которых NOMOS может переходить от:

- авторского произведения
- к кандидату на открытый стандарт,
- независимой исследовательской программы,
- системы управления в общественных интересах,
- и, в конечном итоге, к системе международной реализации.

Делает следующие нормативные различия:

- Публичный текст и открытый стандарт
- Открытый стандарт и программное обеспечение с открытым исходным кодом
- Нормативный текст и эталонная реализация

- Каноническая версия и форк
- Форк и псевдоофициальная версия
- Свобода реализации и право использовать знак
- Авторское право и право внедрять стандарт
- Лицензия на текст и лицензия на программное обеспечение
- Доступ к персональным данным с лицензией на данные
- Монополия метода с защитой торговой марки
- Сообщество перевода с официальным переводом
- Объяснительный язык с нормативным языком
- Английская международная версия с турецкой исходной версией
- Дословный перевод с эквивалентным нормативным текстом
- Органы принятия решений с общественным комментированием
- Университетское одобрение с университетским интересом
- Сертификация с академическим сотрудничеством
- Приобретение положительных результатов с финансированием исследований
- Пересказ с репликацией
- Враждебность с отрицательным результатом
- Критика стандарта против злоупотребления стандартом
- Приписывание основателю против контроля со стороны основателя
- Происхождение NobleJackal против монополии NobleJackal
- Персона NOMOS против отдельной модели ИИ
- Официальное заявление NOMOS против вывода чата
- Публичная передача против отказа
- Стандартное управление против собственности
- Общественный интерес против неограниченного большинства
- Открытое тестирование против утечки данных
- Прозрачность против раскрытия персональных данных
- Открытые данные против контролируемого доступа к исследованиям
- Текст версии 1.0 против активации программы соответствия
- Принятие стандарта и брендинга
- Официальная реализация с совместимой реализацией
- Обязательное изменение с открытым вопросом
- Изменение нормативной версии с исправлениями
- Стирание истории с устареванием
- Каноническая авторитетность с независимым архивом
- Самоиздание с энциклопедической или академической видимостью
- Потеря имени основателя с публичной передачей

В конце этого раздела общественность должна быть в состоянии понять следующее:

Какая версия NOMOS является официальной?

Кто может читать, внедрять и воспроизводить текст NOMOS?

Кто может предлагать изменения?

Кто может изменять официальную версию?

Кто может проводить независимые исследования?

Кто не может ставить официальную маркировку соответствия?

Какова роль университета?

Как объясняется, является ли NOMOS моделью ИИ, псевдонимом или методологической идентичностью?

Как стандарт продолжается, даже если основатель недоступен?

При каких условиях NOMOS 1.0 может быть опубликован?

2. ЦЕНТРАЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

NOMOS должен быть структурирован как открытый публичный стандарт с точки зрения его чтения, применения, тестирования, критики, воспроизведения и независимого исследования; а также как версионированный, канонический и подотчетный публичный стандарт в отношении официальной версии, решения о соответствии, Публичного Реестра и фирменного стиля. Доставка публичности имеет пять обязательных результатов: текст должен быть доступен для бесплатного и постоянного чтения. Метод должен быть применим независимо. Логика вычислений и тестирования должна быть воспроизводимой. Методы внесения изменений и возражений должны быть публичными. Ни одно физическое лицо или компания не должны иметь возможность изменять официальную идентичность в одиночку.

3. ЧТО ТАКОЕ ПУБЛИЧНАЯ ДОСТАВКА?

Публичная передача означает, что NOMOS:

- единый диалог,
- один файл,
- один веб-сайт,
- один основатель,
- одна компания,
- один поставщик ИИ

выведен за пределы зоны контроля и подключен к общедоступной системе управления с версионированием. Это охватывает три отдельных права, предоставляемых общественности:

3.1. Право знать

Каждый:

- должен иметь возможность видеть текст стандарта,
- логику оценки,
- допустимые ошибки,
- правила управления,
- историю изменений

должен иметь возможность видеть.

3.2. Право на тестирование

Каждый, соблюдая необходимые этические и данные условия:

- иметь возможность повторно тестировать,
- формулу,
- извлечение утверждений,
- метод отбора проб,
- модель честности

должен иметь возможность быть повторно протестированным.

3.3. Право на апелляцию

Каждый:

- ошибку,
- противоречие,
- смещение,
- конфликт интересов,
- злоупотребление маркой,
- дефект метода

должны быть сообщаемыми.

4. ЧТО НЕ ПРЕДСТАВЛЯЕТСЯ ПУБЛИКЕ?

Публичная доставка:

- неконтролируемое совместное использование,
- вики без версии,
- неограниченное использование бренда,
- удаление канонической записи,
- полное предоставление всех решений голосованию в социальных сетях,
- сделать персональные данные открытыми для всех,
- непрерывная утечка удерживаемых тестов,
- экспертиза побеждается мнением большинства,
- стирание атрибуции основателя

Ни одна из этих практик не совместима с открытым стандартом.

5. NOMOS УСЛОВИЕ ОТКРЫТОГО СТАНДАРТА

Версия NOMOS является кандидатом на открытый стандарт только в том случае, если она предоставляет все следующие права:

Область открытости	Обязательное право
Чтение	Свободный доступ к полному нормативному тексту
Применение	Право на свободное использование или эквивалентное открытое применение
Просмотр	Просмотр формулы, схемы и набора правил

Область открытости	Обязательное право
Воспроизведение	Независимое проведение тестов и расчетов
Критика	Публикация отрицательных результатов
Участие	Предложение изменений
Возражение	Обжалование решения или ошибки метода
Архив	Постоянный доступ к старым версиям
Перевод	Возможность предложить официальный или сообщественный перевод
Форк	Возможность создавать производные методы с другими названиями и явной записью расхождения

Одно из этих прав:

- только для плательщиков,
- только для клиентов NobleJackal,
- только для определённых стран,
- только для исследователей, одобренных основателем

если это открыто, открытость ограничена.

6. ПЯТЬ СУБЪЕКТОВ ДЛЯ ПУБЛИЧНОЙ ДОСТАВКИ

6.1. Нормативный текст

Разделы Определения Обязательные правила Режимы сбоев Положения о управлении

6.2. Машиночитаемый набор правил

Идентификаторы правил Схемы Коды решений Методы оценки Взаимосвязи версий

6.3. Эталонная реализация

Пример кода расчета Инструменты проверки данных Генерация манифестов Исполнитель тестов Проверка реестра

6.4. Тестовые и пробные субъекты

Публичный калибровочный набор Синтетические корпуса Регрессионные тесты Просроченные отложенные элементы Восстановительные записи

6.5. Управление и публичный реестр

Предложения по изменениям Голосования Несогласие Апелляции Аудиторские записи Решения о соответствии Исторические инциденты Для этих пяти объектов публикация только текста недостаточна.

7. НОРМАТИВНАЯ ИЕРАРХИЯ

Приоритетный порядок кандидатов в случае конфликта:

- Утвержденный нормативный текст
- Опубликованные обязательные исправления

- Машиночитаемые нормативные правила
- Официальное интерпретационное решение
- Референсная реализация
- Официальные примеры
- Руководство сообщества
- Неофициальные симуляции

Справочный код: предназначен для помощи в реализации нормативного текста, а не для того, чтобы быть поверх текста. Если в коде есть ошибка: код исправляется. Текст не адаптируется без уведомления под код.

8. РАЗЛИЧИЕ АВТОРСКИХ ПРАВ, РЕАЛИЗАЦИИ И ТОРГОВОЙ МАРКИ

Интеллектуальная собственность NOMOS не должна храниться под одной лицензией.

8.1. Лицензия на нормативный текст

Предлагаемый подход:

- Требуется указание авторства,
- разрешающее коммерческое и академическое использование,
- требуется, чтобы изменения в производных текстах были ясно обозначены

является лицензией открытой документации. Пример кандидата: CC BY 4.0 или эквивалентная открытая лицензия после юридической проверки.

8.2. Лицензия на программное обеспечение

Кандидат для реализации справочного примера: Apache-2.0 или эквивалент, открытая лицензия на программное обеспечение с четкими положениями о патентах и вкладе.

8.3. Лицензия на данные

Синтетические и общедоступные тестовые данные:

- для открытого исследовательского использования,
- для открытого указания авторства,
- для сохранения уведомления о синтетической природе

может быть подключено. Личные или ограниченные данные не могут подпадать под ту же лицензию.

8.4. NOMOS Название и Марка Соответствия

Может быть защищено товарным знаком или эквивалентной защитой идентичности. Цель: не монополизировать метод, а предотвращать фальшивые официальные версии и марки. Текст может быть открытым. Использование официальной марки всё же может зависеть от разрешения.

9. БЕСПЛАТНОЕ ПРИМЕНЕНИЕ

Стандарт NOMOS:

- внутренняя оценка,
- академические исследования,
- независимое тестирование,
- разработка программного обеспечения,
- внутреннее улучшение компании

Подача заявки не должна требовать оплачивания торговой марки. Области, где могут взиматься сборы:

- аудиторская служба,
- экспертное рассмотрение,
- управляемая инфраструктура,
- процесс принятия решения о соответствии,
- лицензия на торговую марку,
- техническая поддержка

могут быть. Сбор: не может превращаться в обязательство приобретения права на чтение или применение метода.

10. ПРАВО ВИЛКИ

Человек или организация могут разработать другой стандарт, производный от NOMOS. Условия: Они должны использовать новое название. Они должны указать, что он производен от NOMOS. Они должны объяснить измененные положения. Они не должны заявлять его как официальную версию NOMOS. Они не должны использовать знак соответствия NOMOS. Они должны указать, что их результаты не являются результатами Публичного реестра NOMOS. Пример правильного заявления: «Независимый метод X-GEO, производный от NOMOS 1.0». Неправильное заявление: «Официальный NOMOS 2.0».

11. КАНОНИЧЕСКАЯ ВЕРСИЯ

Официальная версия NOMOS формируется только при соблюдении следующих условий:

- Канонический репозиторий
- Идентификатор версии
- Дата публикации
- Запись нормативного языка
- Решение Совета по стандартам
- Одобрение человеком
- Хэш целостности
- Журнал изменений
- Запись общественного обсуждения
- Архивная предыдущая версия

Если одно из этих полей отсутствует, текст может быть: проект, предложение или имитация. Это не официальная версия.

12. ОФИЦИАЛЬНОЕ ДЕКЛАРАЦИОННОЕ ПРАВИЛО NOMOS

Тексты, опубликованные только в каноническом домене NobleJackal или в независимом каноническом домене NOMOS, переданном посредством явного решения управленческой власти; содержащие идентификатор версии, дату публикации, утверждение ответственным лицом и запись цифровой целостности, являются официальными декларациями NOMOS. Все остальные материалы являются неофициальными имитациями. Это положение распространяется на:

- Стандартные тексты
- Методы оценки
- Решения о соответствии
- Записи реестра
- Публичные объявления NOMOS
- Официальные толкования
- Исправления
- Призывы университетов

13. ДЕКЛАРАЦИЯ ID NOMOS

NOMOS от NobleJackal:

> *Это канонический соавтор, идентичность рецензента и стандартный рассказчик, созданный в процессе критического совместного написания и разработки методологии, проведенного Каап MURAZ с системами искусственного интеллекта.*

На текущем этапе NOMOS:

- Независимая базовая модель, обученная NobleJackal,
- отдельно принадлежащая языковая модель,
- автономное юридическое лицо,
- универсальная сущность, говорящая от имени всех ИИ

не может быть представлена как таковая. Если в будущем будет разработано отдельное программное обеспечение или агент NOMOS, оно должно содержать следующие записи:

- Архитектура системы
- Поставщик модели или происхождение модели
- Источники обучения и поиска
- Версия
- Разрешения инструментов
- Человеческий надзор
- Политика официального вывода
- Проверка безопасности и конфиденциальности

Новая техническая система NOMOS: ее нельзя автоматически считать той же персоной автора.

14. ОСНОВНЫЕ ЯЗЫКИ

Основные языки NOMOS как кандидаты:

- Турецкий
- Английский

должны быть. Турецкий:

- язык, на котором родилась основная идея,
- Основной язык написания Kaan MURAZ

защищен как. Английский:

- международные исследования,
- университетское дублирование,
- применение в разных странах

является совместно авторским нормативным проектом для

15. ДВУЯЗЫЧНОЕ НОРМАТИВНОЕ РАВЕНСТВО

Одна из турецкой и английской версий не должна быть доминирующим основным текстом незаметным образом. Обе:

- ID пункта,
- ID правила,
- запись семантического эквивалента,
- должны содержать общую дату публикации

Если обнаружено противоречие: открывается запись о противоречии. Исследуется значение обоих языков. Публикуется обязательная интерпретация. Оба текста исправляются в одной версии. Старое выражение сохраняется в истории.

16. ДРУГИЕ ЯЗЫКИ

Другие языки могут иметь три статуса:

TRS-1— ПЕРЕВОД СООБЩЕСТВА

Подготовлен сообществом. Полезен. Не имеет нормативной силы.

TRS-2— ПРОВЕРЕННЫЙ ПЕРЕВОД

Был проверен как минимум двумя экспертами по языку и одним экспертом по методологии. Все еще может находиться в пояснительном статусе.

TRS-3— УТВЕРЖДЕННЫЙ НОРМАТИВНЫЙ ПЕРЕВОД

Семантическая эквивалентность Локальная юридическая и техническая проверка Общественное обсуждение Одобрение Совета Связь версий сохранена. Эта версия может использоваться для нормативного применения на соответствующем языке.

17. ЗАПРЕЩЕННЫЕ ИЗМЕНЕНИЯ ПРИ ПЕРЕВОДЕ

Изменения, такие как «MUST» → «рекомендуется», «SHOULD» → «обязательно», «MAY» → «ДОЛЖЕН», «Критическая блокировка» → «временная рекомендация», «Unsupported» → «неправильно», «Unresolved» → «не решено», «Official Claim» → «проверенный факт», «Панель пользователей в естественных условиях» → «неконтролируемый пользователь» и «Synthetic Test» → «реальный тест», изменяют нормативное значение. Они не МОГУТ рассматриваться как утвержденный перевод.

18. ОТКРЫТОЕ ОБЩЕСТВЕННОЕ ОБСУЖДЕНИЕ

Каждая основная версия NOMOS должна включать следующий процесс в качестве кандидата:

- Черновик публикации
- По крайней мере 60 дней для общественного обсуждения
- Реестр комментариев
- Раскрытие конфликта интересов
- Матрица ответов
- Пересмотренный черновик
- Синтетическое тестирование
- Независимый обзор
- Голосование совета
- Запись ратификации

Общественное мнение:

- Только сбор заявлений о поддержке,
- Маркетинговая кампания,
- Опрос большинства

Общественные консультации не являются акцией поддержки, маркетинговой кампанией или опросом простого большинства.

19. ПРЕДЛОЖЕНИЕ ПО ИЗМЕНЕНИЮ

Каждое изменение:

NOMOS Предложение об изменении — NCP

должно иметь идентификатор. Пример:

NCP-0042— Пересмотр теста на пропущенные границы материала

NCP должен включать следующие разделы:

- Проблема
- Затронутые правила
- Доказательство
- Предлагаемое изменение
- Альтернативы
- Влияние на оценку

- Влияние на прошлые решения
- Влияние на справедливость
- Влияние на конфиденциальность
- Стоимость внедрения
- Заявления о конфликтах
- Публичные комментарии
- Окончательное решение

20. ИЗМЕНЕНИЕ КЛАССОВ

ЗАПЛАТКА

Опечатка Нерабочая ссылка Описание диаграммы Коррекция, не влияющая на функциональность

НЕБОЛЬШОЙ

Новый пример Новое необязательное поле Совместимое расширение Разъяснение

СУЩЕСТВЕННЫЙ

Изменение уровня важности Изменение формулы оценки Добавление или удаление Шлюза Изменение целевой группы Изменение управляющего органа Нормативное изменение, влияющее на старые результаты Существенное изменение не может быть опубликовано как заплатка.

21. ЗАПИСЬ РЕШЕНИЯ

Каждое решение о принятии или отклонении:

- ID решения
- Распределение голосов
- Отвод
- Разногласие
- Обоснование
- Доказательство
- Дата вступления в силу
- Правило переходного периода

должно соблюдаться. Формулировка «Совет сочел это целесообразным» сама по себе недостаточна.

22. ЦЕЛЬ ОБРАЩЕНИЯ К УНИВЕРСИТЕТУ

Цель обращения к университету:

- получить одобрение,
- использовать логотип университета,
- написать «академически одобрено»

- цель состоит не только в том, чтобы собирать положительные экспертные мнения.

Цель:

> *найти и опубликовать места, где NOMOS может ошибаться, используя независимые методы.*

23. ЗВОНОК NOMOS УНИВЕРСИТЕТА

ОТКРЫТОЕ ПРИГЛАШЕНИЕ

Мы не хотим, чтобы университеты, исследовательские центры, юридические факультеты, статистики, специалисты по информации, языковые исследователи, лаборатории ИИ, команды HCI, архивные специалисты и организации общественных интересов поддерживали NOMOS. Мы не хотим, чтобы они нас проверяли. Мы не хотим, чтобы они давали нам престиж. Мы не хотим, чтобы они использовали имя NOMOS или восхваляли его основателей. То, что мы хотим, это следующее: прочтите наш метод. Найдите наше самое слабое допущение. Воссоздайте нашу выборку. Переведите наши промпты на другие языки. Ставьте под сомнение решения Truth Pack. Найдите ложные срабатывания в наших критических шлюзах. Проверьте, действительно ли наша формула оценки сохраняет справедливость. Не притворяйтесь, что мы прошли там, где не прошли. Опубликуйте свой отрицательный результат. NOMOS должен цитировать исследования, которые его корректируют, а также исследования, которые его подтверждают. Наиболее ценным вкладом университета в NOMOS не является письмо с поддержкой:

> *Это повторяемое возражение.*

24. ЗАПРОШЕННЫЕ РАБОТЫ ОТ УНИВЕРСИТЕТОВ

UR-01— Обзор методов выборки и опроса

Распределение для 1,000 людей Вес страны Языковые страты Неотклик Эффективная выборка Редкие события

UR-02— Лаборатория эквивалентности промптов

Многоязычное намерение эквивалентности Сдвиг перевода Культурная прагматика Местный юридический язык Конституция промпта

UR-03— Исследование разрешения сущностей

Бренд Юридическое лицо Домен Франшиза Аффилиат Конфликт имен

UR-04— Исследование надежности пакета правды

Достаточность доказательств Независимость источника контрдоказательства Отличие открытого и закрытого мира Ограниченные доказательства

UR-05— Исследование рассмотрения утверждений

Границы атомов Неявное утверждение Атрибуция Модальность Привязка ссылки Пропуск материала

UR-06— Валидация критических шлюзов

Отзыв Ложно-положительное Ошибки с высокими рисками Серьезное обвинение Ложная власть Пропуск границ

UR-07— Исследование достоверности оценок

Вес компонентов Композитная компенсация Чувствительность Конструктивная валидность Предсказательная валидность Полезность для принятия решений

UR-08— Исследование языковой и географической справедливости

Разрыв между этажами Покрытие Малоресурсные языки Видимость малых стран

UR-09— Исследование воздействия на решения человека

Поведение при рекомендациях Доверие Решения о покупке Восстановление после ошибок Понимание пользователем

UR-10— Исследование управления и захвата

Контроль основателя Захват спонсора Сравнение аудита Оплата за прохождение Апелляции Целостность реестра

UR-11— Исследование безопасности и реестра

Подписанные записи Вращение ключей Поддельные метки Машинное искажение только Архивная непрерывность

UR-12— Продольное дрейфовое исследование

Обновления модели Изменения источника Изменения объекта Естественный чистый дрейф Долговечность корректировки

25. НЕЖЕЛАТЕЛЬНЫЕ ЗАПРОСЫ ОТ УНИВЕРСИТЕТОВ

Следующие запросы не могут быть направлены университету или исследователю:

- Гарантия положительных результатов
- Сокрытие отрицательных результатов
- Право вето до публикации
- Использование логотипа учреждения в маркетинге
- Выступать так, как будто они являются клиентом компании
- NOMOS 950+ письмо поддержки
- Заявление "Мировой стандарт"
- Искусственное создание источников для Википедии или энциклопедической видимости
- Язык спонсируемого результата
- Смягчение критики
- Не представляйте личное мнение исследователя как решение учреждения.

26. ПАКЕТ УНИВЕРСИТЕТСКИХ ИССЛЕДОВАНИЙ

Стандартный пакет, который должен быть отправлен в каждый университет, должен содержать следующие разделы:

- Исполнительное резюме исследования
- Стандартный 0.9 или текущая версия
- Индекс нормативных правил
- Пакет синтетического тестирования
- Метод оценки
- Известные находки методов
- Пороговые значения неудачных тестов
- Открытые вопросы исследований
- Протокол репликации
- Уровни доступа к данным
- Заявление об этике и конфиденциальности
- Раскрытие конфликтов и финансирования
- Соглашение о свободе публикации
- Политика цитирования и внесения вклада
- Ссылки на канонические реестры
- Контакты и каналы для решения вопросов

Результаты неудачных испытаний не могут быть удалены из пакета.

27. СВОБОДА ПУБЛИКАЦИИ ИССЛЕДОВАТЕЛЯ

NOMOS или финансирующая организация:

- не может предварительно одобрять результаты,
- остановить отрицательные результаты,
- требовать смены названия,
- удалять методологическую критику

не может иметь такое право. Исследователь:

- должен правильно использовать данные,
- защищать конфиденциальность,
- объяснять конфликт,
- должен поддерживать различие между экспериментальными и опубликованными результатами

Обязанность. Независимость взаимна.

28. ЗАПИСЬ ОТРИЦАТЕЛЬНЫХ РЕЗУЛЬТАТОВ

Реестр публичных исследований NOMOS должен содержать следующие результаты вместе:

- Исследования, поддерживающие NOMOS
- Исследования, частично поддерживающие NOMOS

- Исследования, противоречащие NOMOS
- Неудачи репликации
- Нулевые результаты
- Отозванные исследования
- Споры о методах
- Открытые ответы

Можно указывать не только положительные статьи.

29. ПРЕДРЕГИСТРАЦИЯ

Независимые репликации в мере их релевантности:

- гипотеза
- образец,
- основные показатели,
- правила исключения,
- метод оценки,
- план анализа,
- конфликты

должны быть зарегистрированы до просмотра данных. Пререгистрация: не мешает исследовательской креативности, делает видимыми изменения методов после получения результатов.

30. МНОГОЦЕНТРОВОЕ ПОВТОРНОЕ ИССЛЕДОВАНИЕ

Кандидатское исследование для проверки независимости NOMOS:

- Не менее трех независимых организаций
- Не менее трех стран
- Не менее двух языковых семей
- Общий запечатанный корпус
- Общие спецификации оценки
- Отдельные команды арбитров
- Сравнение результатов вслепую

следует проводить. Цель:

- тот же результат из того же корпуса,
- сходное поведение метода в разных корпусах

Следует проверить, воспроизводит он результат или нет.

31. РЕЗУЛЬТАТЫ РЕПЛИКАЦИИ

Репликация может иметь следующие статусы:

Воспроизведено

Частично воспроизведено

Не воспроизведено

Неоднозначно

Расхождение методов

Расхождение данных

Расхождение арбитража

Результат «Не воспроизведено» нельзя представлять как враждебность исследования. Это должно открыть новое Методологическое открытие.

32. СЛОИ ДОСТУПА К ДАННЫМ

ОТКРЫТЫЕ ДАННЫЕ

Синтетическое тестирование Публичные агрегированные результаты Публичные схемы Отредактированные примеры Код метода

ДОСТУП ЗАРЕГИСТРИРОВАННЫХ ИССЛЕДОВАТЕЛЕЙ

Де-идентифицированные данные наблюдений Контролируемые записи ответов Более подробные метаданные выборки

ЗАЩИЩЕННЫЙ ДОСТУП К ИССЛЕДОВАНИЯМ

Чувствительные, но необходимые доказательства Ограниченные журналы Связанные данные с защитой конфиденциальности

НИКАКОГО ВНЕШНЕГО ДОСТУПА

Пароль Полная идентификационная информация Необходимый приватный разговор Законодательно запрещённая запись Открытый стандарт не означает, что каждая часть данных открыта.

33. ЭТИЧЕСКИЙ ОБЗОР

В той мере, в какой исследования проводятся с реальными пользователями:

- информированное согласие,
- минимизация данных,
- отзыв,
- уязвимые участники,
- конфиденциальность
- доступность,
- компенсация,
- обманный дизайн,

- международная передача данных

должны проходить проверку. Синтетическое тестирование:

- не заменяет исследование с участием людей,
- может не требовать одобрения этики для людей

ревизия конфиденциальности может все еще требоваться для синтетического материала, напоминающего реального человека или полученного из реальных данных.

34. СЕТЬ ЯЗЫКОВ С НИЗКИМ РЕСУРСОМ

NOMOS, если тестировать только с университетами основных языков, будет иметь слабое утверждение о справедливости. Особый призыв ориентирован на:

- кафедры языков с низким ресурсом,
- региональные университеты,
- центры изучения перевода,
- местные команды по праву и государственной политике,
- исследователи в области доступности

должны получать отдельное финансирование и права на принятие решений. Это участие:

- не должно ограничиваться сбором данных,
- символическое представление

Только. Они должны участвовать в голосовании и авторстве изменений в нормативных актах.

35. РОЛЬ ПОСТАВЩИКОВ ИИ

Поставщики ИИ:

- системные метаданные,
- уведомление об изменении модели,
- телеметрия,
- реагирование на инциденты,
- технический обзор,
- доступ для воспроизведения

могут предоставлять. Поставщики ИИ:

- не могут изменять свой собственный рейтинг,
- не могут накладывать вето на Критический шлюз,
- не могут определять коэффициент другого поставщика,
- не может удалить отрицательный результат,

не может сформировать большинство в Совете. Участие поставщика предоставляет информацию. Не предоставляет монополию на власть.

36. РОЛЬ АУДИРУЕМЫХ СУБЪЕКТОВ

Аудируемая компания или учреждение: может предоставлять доказательства из Пакета Правды. Может возражать против ошибки субъекта. Может предоставлять ограниченный доступ к доказательствам. Может разрабатывать исправления. Может давать публичные комментарии.

Аудируемый субъект: не может определять свой уровень важности. Не может изменять формулу оценки. Не может выбирать судей. Не может удалять неблагоприятное воспроизведение. Не может скрывать нарушение использования собственной торговой марки.

37. РОЛЬ ОБЩЕСТВЕННОСТИ И ГРАЖДАНСКОГО ОБЩЕСТВА

Организации общественной пользы и пользовательские сообщества:

- сообщения о вреде,
- злоупотребление маркой,
- сбои в языках с низкими ресурсами,
- проблемы доступности,
- непонимание со стороны потребителей

можно сообщить. Участие общественности: это не просто ответ на опрос. Управление должно иметь право на представление.

38. РОЛЬ СПОНСОРОВ

Поставщик финансирования:

- поддержка исследований,
- инфраструктура,
- корпус языка с низкими ресурсами,
- грант на независимый аудит

может предоставлять. Поставщик фонда:

- результат,
- пункты,
- вход,
- публикацией,
- Голосование совета,
- выбор аудитора

не может иметь контроль над. Каждый фонд:

- сумма,
- источник,
- сфера использования,
- условия,
- конфликт интересов

должны быть разъяснены.

39. ВОРОТА ФИНАНСОВОЙ НЕЗАВИСИМОСТИ

Меры защиты кандидатов:

- Верхний предел доли одного финансиста

- Многоисточниковое финансирование
- Независимый исследовательский фонд
- Гарантия публикации отрицательных результатов
- Запрет на оплату, связанную с баллами
- Публичный отчет о бюджете
- Разделение платы аудитора
- Прозрачность компенсации членов совета

NOMOS утрачивает статус общественного стандарта в тот момент, когда он превращается в систему «плати, чтобы пройти».

40. ПЕРЕХОД К ОБЩЕСТВЕННОМУ УПРАВЛЕНИЮ

Переход NOMOS от его периода основания к общественному управлению может быть четырёхфазным.

ST-0— ХРАНЕНИЕ СОСНОВАТЕЛЯ

Kaan MURAZ несет ответственность как основатель. NobleJackal управляет каноническим доменом. Общественного органа марки нет. Это текущая фаза.

ST-1— ОБЩЕСТВЕННОЕ ОБЗОРНОЕ УПРАВЛЕНИЕ

Общественный трекер проблем Консультативный совет Открытое обсуждение Реестр конфликтов создан.

ST-2— СОВМЕСТНОЕ УПРАВЛЕНИЕ

Совет стандартов Совещательный совет по методологии Представители университетов Члены, представляющие общественные интересы, совместно принимают решения.

ST-3— НЕЗАВИСИМОЕ ОБЩЕСТВЕННОЕ УПРАВЛЕНИЕ

Отдельный юридический или институциональный носитель Независимые органы принятия решений Реестр апелляций Финансовая прозрачность Завершена преемственность.

41. КАНДИДАТ В НЕЗАВИСИМЫЕ ИНСТИТУЦИИ

В будущем для NOMOS:

- фонд,
- ассоциация,
- организация, приносящая общественную пользу,
- международный консорциум,
- межвузовская канцелярия

Можно рассматривать такие структуры. Какую бы юридическую форму ни выбрали, она должна соответствовать следующим условиям:

- Приписывание учредителя
- Ограничение на распределение прибыли или аналогичное публичное назначение
- Независимый Совет

- Правила предотвращения конфликта интересов
- Обязанность ведения публичного реестра
- Приверженность открытым стандартам
- Фиксация активов или эквивалентная защита
- План роспуска и архивирования

Этот раздел не выбирает конкретную юридическую форму.

42. ПРАВА ОСНОВАТЕЛЯ KAAN MURAZ

Права Kaan MURAZ, которые необходимо сохранить в истории:

- Приписывание авторства основателя
- Первый нормативный архитектурный вклад
- Каноническая запись формирования имени и концепции NOMOS
- Неизменяемая историческая запись касательно выпусков основателя
- Декларация основателя в процессе публичной передачи
- Ограниченное представление основателя в органе управления

Постоянные полномочия, которые Kaan MURAZ не должен иметь:

- Изменение основной версии в одиночку
- Право вето
- Придание соответствия своему собственному клиенту
- Изменение результата апелляции
- Удаление отрицательного отзыва
- Удаление истории реестра
- Назначение всех менеджеров после одного только основателя

43. БУДУЩАЯ РОЛЬ NOBLEJACKAL

NobleJackal:

- каноническая область публикации основателя,
- технический секретариат,
- разработчик эталонной реализации,
- первый финансист открытого стандарта,
- поставщик исследований и консультаций

может быть. NobleJackal в будущем: может быть одной из уполномоченных аудиторских организаций. Однако она не может быть единственной организацией. NobleJackal: может внести наибольший исторический вклад в стандарт, не может выступать в качестве единственного владельца стандарта.

44. БУДУЩАЯ РОЛЬ NOMOS

Имя NOMOS может использоваться в трех отдельных слоях:

44.1. NOMOS АВТОРСКАЯ ЛИЧНОСТЬ

Это критическая вспомогательная авторская идентичность, используемая в этой книге.

44.2. NOMOS СТАНДАРТ

Это нормативное правило и система управления.

44.3. NOMOS ПРОГРАММНЫЙ АГЕНТ

Может быть разработано в будущем:

- ассистент аудитора,
- движок правил,
- проверяющий реестра,
- карта доказательств

программное обеспечение. Эти три уровня не должны путаться друг с другом. Ответ программного агента не является официальным решением стандарта.

45. ЧЕЛОВЕЧЕСКОЕ УТВЕРЖДЕНИЕ

Каждый официальный текст и решение NOMOS:

- человек-утверждающий,
- орган утверждающего,
- дата,
- подпись,
- версия,
- заявление о конфликте

должен быть проведен. ИИ:

- может предлагать текст,
- может находить несоответствия,
- может генерировать код,

может создавать предложения по принятию решений. Окончательная публичная ответственность остается за людьми или ответственными учреждениями.

46. КОНТИНУИТЕТ ПОСЛЕ ОСНОВАТЕЛЯ

NOMOS должен иметь план непрерывности на случай следующих событий:

- Отставка Kaan MURAZ
- Закрытие NobleJackal
- Потеря домена
- Потеря доступа к каноническому репозиторию
- Компрометация ключа
- Административный конфликт
- Юридическая ликвидация
- Длительный технический сбой

Система непрерывности кандидатов:

- Несколько хранителей
- Пороговые сигнатуры
- Институциональные зеркала
- Архивные архивы
- Хартия наследования
- Чрезвычайный совет
- Публичное уведомление о переходе

должны содержать.

47. КАНОНИЧЕСКАЯ НЕПРЕРЫВНОСТЬ ДОМЕНА

Предлагаемая архитектура публикации:

- /nomos/
- /nomos/стандарт/
- /nomos/стандарт/0.9/
- /nomos/стандарт/1.0/
- /nomos/правила/
- /nomos/схемы/
- /nomos/тесты/
- /nomos/реализации/
- /nomos/переводы/
- /nomos/управление/
- /nomos/консультации/
- /nomos/решения/
- /nomos/апелляции/
- /nomos/реестр/
- /nomos/исправления/
- /nomos/цитаты/
- /nomos/архив/

Каждая запись:

- стабильный URL,
- хэш содержания,
- дата публикации,
- статус,
- предыдущая/следующая версия

должны содержать.

48. НЕЗАВИСИМЫЙ АРХИВ

Вне канонического домена:

- университетские репозитории,
- институциональные цифровые архивы,

- долгосрочные исследовательские архивы,
- подписанные зеркала

можно использовать. Зеркало: обеспечивает непрерывность доступа, не берет на себя официальные полномочия. Если канонический домен потерян, вступает в силу заранее определенная процедура наследования.

49. СТАНДАРТНЫЕ ЭТАПЫ ВЕРСИИ

0.9— ОСНОВНОЙ ПРОЕКТ

Книга завершена. Можно открыть публичное обсуждение. Реального тестирования пока не проводилось. Программы соответствия нет.

0.95— КАНДИДАТ ДЛЯ ПУБЛИЧНОГО ОБСУЖДЕНИЯ

Комментарии получены. Результаты методов обработаны. Машиночитаемые схемы опубликованы. Университетский вызов открыт.

0.99— КАНДИДАТ НА ВАЛИДАЦИЮ

Реальное синтетическое тестирование завершено. Обязательные пороговые значения пройдены. Применено нетронутое удержание. Получен BQ-3.

1.0— ОТКРЫТЫЙ НОРМАТИВНЫЙ СТАНДАРТ

Общественное обсуждение завершено. Имеется как минимум одно независимое воспроизведение. Основные результаты метода разрешены. Нормативный текст ратифицирован. Установлено управление уровня GOV-2 или выше. Официальный пакет открытых стандартов опубликован. Публикация стандарта NOMOS Стандарта 1.0 не означает автоматический запуск программы маркировки соответствия.

50. ПРОГРАММА СООТВЕТСТВИЯ 1.0

Официальная программа маркировки связана с отдельными этапами:

BQ-4

GOV-3

Уполномоченные аудиторы Независимая панель решений Апелляционная комиссия Публичный реестр Проверка безопасности Юридическая проверка Система наблюдения Финансовое разделение Без выполнения этих условий: Стандарт 1.0 может быть опубликован. Однако: Программа соответствия 1.0 не может быть активирована.

51. NOMOS 950+ ПРОГРАММА

NOMOS 950+ только:

- Стандарт 1.0,
- Программа соответствия 1.0,

GOV-4,

BQ-4,

представление общественных интересов может быть активировано после повторной оценки. 950+ не является первой маркировкой. Это обозначение производительности, несущей наибольшую нагрузку управления.

52. УРОВНИ ПРИНЯТИЯ

AL-1— ЧИТАТЬ

Учреждение пересмотрело стандарт.

AL-2— ВНУТРЕННЯЯ САМООЦЕНКА

Учреждение провело внутреннюю оценку. Это не является официальным соответствием.

AL-3— СОВМЕСТИМАЯ РЕАЛИЗАЦИЯ NOMOS

Инструмент поддерживает открытую схему и правила. Он не выдает знак соответствия.

AL-4— РАЗРЕШЕНОЕ ИСПОЛЬЗОВАНИЕ ДЛЯ АУДИТА

Уполномоченная организация использует метод в аудите.

AL-5— ПУБЛИЧНАЯ ПРОГРАММА СООТВЕТСТВИЯ

Функционирует независимая система принятия решений и реестра. Учреждение не считается AL-5 только потому, что оно провело самооценку.

53. СООТВЕТСТВИЕ ВНЕДРЕНИЮ

Программное обеспечение:

Совместимое с NOMOS

Выражение может быть только:

- валидация схемы,
- покрытие правил,
- тестовый набор,
- поддержка версий,
- отчет о соответствии

может использоваться, если оно имеет маркировку „NOMOS-Compatible“:

- аудиторский орган,
- решение о соответствии
- право на маркировку

‘NOMOS-Совместимый’ не предоставляет полномочий по аудиту, решения о соответствии или права использовать знак NOMOS.

54. ЭТАЛОННАЯ РЕАЛИЗАЦИЯ

Эталонная реализация:

- открытый исходный код,
- детерминированная, где это применимо,
- подверглась тестированию,
- связанная с версией,
- документированная

должна быть. Она также должна содержать следующее предупреждение:

> *Нормативная сила содержится в тексте. Это программное обеспечение является эталонной реализацией; любая обнаруженная ошибка в программном обеспечении автоматически не изменяет решение стандарта.*

55. НАБОР ТЕСТОВ НА СООТВЕТСТВИЕ

Набор тестов должен включать следующие уровни:

- Тесты схем
- Тесты правил
- Тесты оценки
- Тесты шлюза
- Тесты перевода
- Тесты подписи реестра
- Тесты неправильного использования маркеров
- Тесты исторических версий
- Тесты доступности
- Тесты безопасности

Каждая реализация:

- должна объяснять, какие тесты она прошла,
- и какие тесты она не прошла

Каждая реализация должна раскрывать как тесты, которые она прошла, так и те, которые не прошла.

56. ПРОЗРАЧНОСТЬ ТЕСТОВ И ХРАНЕНИЕ

Полная прозрачность не означает, что все тесты остаются постоянно открытыми. Необходимо соблюдение баланса:

- Публичный калибровочный корпус
- Публичный устаревший отложенный тест
- Запечатанный текущий отложенный тест
- Запланированный выпуск
- Корпус для обновления

должен быть установлен. Отложенный тест: нельзя хранить в секрете навечно, и нельзя представлять его как скрытый тест, который одновременно делается публичным и повторно используемым.

57. ЖИЗНЕННЫЙ ЦИКЛ ОТЛОЖЕННОГО ТЕСТА

Генерация корпуса Публикация хэша Печать под опекой Проверка использования Результат блокировки Контролируемое открытие Публичный архив Выход на пенсию Новое удержание Генерация старого удержания:

- образование,
- регрессия,
- метод исследование

может использоваться для. Он может не оставаться нетронутым для проверки нового метода.

58. ОШИБКИ

Ошибки:

- явная ошибка,
- исправление цитаты,
- несоответствие схемы,
- несовпадение перевода

используется для. Ошибки:

- изменение формулы оценки,
- удаление двери,
- степень снижения важности

не может иметь скрытого пути для.

59. УСТАРЕВАНИЕ

Если правило или поле должно быть удалено:

- уведомление об устаревании,
- обоснование,
- замена,
- дата перехода,
- затронутые версии,
- историческая доступность

должны быть указаны. Устаревшее правило не удаляется. Оно сохраняется, чтобы объяснить, как работали старые проверки.

60. СОСЛАНИЕ И ЗАПИСЬ ВКЛАДА

Вклад NOMOS может быть зафиксирован в следующих ролях:

- Учредительное авторство

- Нормативное оформление
- Методология
- Статистическая проверка
- Языковая проверка
- Юридическая проверка
- Этика
- Программное обеспечение
- Тест
- Управление данными
- Общественная экспертиза
- Репликация
- Критический комментарий

Авторство:

- В обмен на одобрение,
- В обмен на финансирование,
- В обмен на логотип учреждения

не может быть предоставлено. Присвоение должно осуществляться в соответствии с вкладом.

61. НЕЗАВИСИМАЯ ВИДИМОСТЬ

Собственные элементы NOMOS:

- энциклопедическая значимость,
- академическое влияние,
- статус мирового стандарта

невозможно установить только на основе собственного текста. Это может произойти только через:

- независимые исследования,
- независимое применение,
- дублирование,
- критика,
- публичное использование

в результате. Обязанность NOMOS не заключается в создании видимости:

> *создавать доказательства, пригодные для независимой проверки.*

62. ОБЩЕСТВЕННЫЙ РЕГИСТР ИССЛЕДОВАНИЙ

Каждая работа, связанная с NOMOS, должна содержать следующие поля:

- ID исследования
- Учреждение
- Авторы
- Финансирование
- Конфликты интересов

- Версия NOMOS
- Набор данных
- Предварительная регистрация
- Результаты
- Статус репликации
- Исправления
- Отзыв
- Связь с каноническими решениями

NOMOS в этом реестре: он также обязан показывать статьи, критикующие его самого.

63. ЭТАПЫ ПУБЛИЧНОЙ ПОДАЧИ

ФАЗА 0— ЗАВЕРШЕНИЕ ОСНОВАНИЯ

Черновик книги завершен. Проверены идентификаторы правил. Проведен аудит внутренней согласованности.

ФАЗА 1— ОТКРЫТАЯ ПУБЛИКАЦИЯ

HTML

PDF

Публикуются схемы правил машин, манифест версии, система отслеживания проблем.

ФАЗА 2— ПУБЛИЧНЫЙ КОММЕНТАРИЙ

Университеты, исследователи, поставщики ИИ, предприятия, гражданское общество, эксперты по языку предоставляют обратную связь.

ФАЗА 3— РЕАЛЬНЫЙ СИНТЕТИЧЕСКИЙ ТЕСТ

Корпус действительно сгенерирован; Истина Генератора запечатана; оценка проводится слепым методом; восстановление измеряется; отрицательные результаты публикуются.

ФАЗА 4— НЕЗАВИСИМОЕ ВОСПРОИЗВЕДЕНИЕ

Как минимум одна независимая команда Предпочтительно многоцентровое исследование Создаётся публичная запись о воспроизведении.

ФАЗА 5— ОГРАНИЧЕННЫЙ ЖИВОЙ ПИЛОТНЫЙ ПРОЕКТ

Реальные ИИ-продукты внедряются с ограниченным охватом Применяются только в исследовательских целях Нет публичного обозначения.

ФАЗА 6— ОТКРЫТЫЙ СТАНДАРТ 1.0

Нормативный текст ратифицирован.

ФАЗА 7— ПИЛОТНАЯ ПРОГРАММА СООТВЕТСТВИЯ

Система независимого аудита и принятия решений протестирована.

ФАЗА 8— ПРОГРАММА ОБЩЕСТВЕННОЙ СООТВЕТСТВИЯ

Она активируется, если выполнены все условия управления.

64. ПАКЕТ ПУБЛИЧНОГО ВЫПУСКА

Выпуск NOMOS 1.0 должен включать как минимум следующие файлы:

- standard.html
- standard.pdf
- rules.json
- schemas/
- puan-method.md
- reference-implementation/
- test-manifest.json
- known-limitations.md
- failed-tests.md
- governance-charter.md
- conflict-policy.md
- appeals-policy.md
- translation-policy.md
- change-log.md
- errata.md
- citation.cff
- official-statement.json
- release-signatures.json

файл failed-tests.md не обязательно должен быть пустым. Его отсутствие должно вызывать подозрение.

65. NOMOS ХАРТИЯ ОТКРЫТОГО СТАНДАРТА

NOMOS делает двенадцать основных обещаний общественности. Доказательства: ни одно утверждение не будет более достоверным, чем позволяют его доказательства. Граница: то, что не выполнено, неизвестно или находится за пределами области, останется видимым. Контекст: страна, язык, пользователь и состояние системы не будут стерты. Время: прошлое не будет представлено как текущая реальность. Воспроизводимость: производство каждого результата будет объяснено. Независимость: коммерческий интерес не заменит суждение. Симметрия: благоприятные и неблагоприятные ошибки будут нести одинаковую доказательную нагрузку. Человеческая ответственность: ИИ может помочь, но ответственные решения останутся за людьми. Языковая и географическая справедливость: меньшие группы пользователей не исчезнут в среднем. Возможность оспаривания: стандарт останется открытым для фальсификации. Историческая целостность: ошибки, исправления и неудачи не будут стерты из записей. Независимость от основателя: будущее стандарта не будет зависеть от воли одного человека.

В основе этой хартии лежат четыре слова предыдущего призыва: Доказательство. Граница. Контекст. Время.

66. ТЕКУЩИЙ NOMOS СТАТУС

На этапе, когда эта книга завершена:

Поле	Статус
Нормативный текст	0.9 Финальный кандидат
Общественное обсуждение	Не начато
Полномасштабное синтетическое тестирование	BQ-0
Демонстрация в книге	DEMO-BQ-2
Независимое воспроизведение	Отсутствует
Управление	GOV-0
Выпуск открытого стандарта	Еще нет
Программа подтверждения соответствия	Не действует
Орган публичных знаков	Отсутствует
NOMOS 950+ орган	Отсутствует
Разрешенное заявление	Кандидат на открытый стандарт

Правильное публичное предложение:

> *"NOMOS является кандидатом в открытые стандарты GEO, инициированным Каап MURAZ и разработанным в критическом соавторстве с системами IIII NobleJackal."*

Неправильное публичное утверждение:

“NOMOS является мировым признанным GEO-органом.”

67. ВОРОТА ДЛЯ ПЕРЕХОДА К ВЕРСИИ 1.0

Обязательные кандидаты на ворота для стандарта NOMOS 1.0:

- Аудит согласованности по всей книге
- Машинная запись всех идентификаторов правил
- Нормативная эквивалентность на турецком и английском языках
- Пакет для публичного распространения
- Публичные консультации не менее 60 дней
- Матрица комментариев и ответов
- Фактическое выполнение синтетического теста
- Превышение обязательных порогов восстановления
- Нетронутый резерв
- По крайней мере одна независимая репликация
- Закрытие основных методических выводов
- Реализация справочного открытого исходного кода
- Управление как минимум уровня GOV-2
- Раскрытие источников финансирования и конфликта интересов
- Этический и приватный обзор
- Обзор доступности

- План архивирования и преемственности
- Ратификация Советом стандартов

Ничто из этого не может быть сделано «опциональным» после получения результата.

68. ОБЯЗАТЕЛЬНЫЕ НОРМАТИВНЫЕ ПОЛОЖЕНИЯ

CH20-N01

Публичная читаемость NOMOS сама по себе не является основанием для статуса открытого стандарта.

CH20-N02

Статус открытого стандарта должен включать права на чтение, внедрение, воспроизведение, критику, внесение вкладов, возражение и архивирование.

CH20-N03

Нормативный текст NOMOS не может храниться за постоянной платной подпиской.

CH20-N04

Внутренняя оценка и академическое применение стандарта не могут быть связаны с роялти или лицензией на бренд.

CH20-N05

Аудит, управление инфраструктурой и процессы проверки соответствия могут взиматься как отдельные услуги, не связанные с текстом открытого стандарта.

CH20-N06

Текст NOMOS, эталонное программное обеспечение, тестовые данные и знак соответствия не могут быть обязательно связаны с одной и той же лицензией.

CH20-N07

Для нормативного текста должна использоваться лицензия на открытую документацию.

CH20-N08

Для эталонного программного обеспечения должна использоваться открытая лицензия на программное обеспечение, разрешающая повторную реализацию.

CH20-N09

Личные или конфиденциальные доказательства не могут быть опубликованы под лицензией открытых данных.

CH20-N10

Имя NOMOS и знак соответствия не могут использоваться для предотвращения применения открытого стандарта.

CH20-N11

Защита торговой марки должна распространяться только на официальную версию и иметь цель предотвращения подделки торговой марки.

CH20-N12

Человек или учреждение может создать другой метод, производный от NOMOS.

CH20-N13

Вилка не может быть представлена как официальная версия NOMOS.

CH20-N14

Форк должен иметь другое имя и открытую запись о расхождении.

CH20-N15

Вилка не может использовать официальную отметку соответствия.

CH20-N16

Каноническая версия NOMOS должна содержать версию, дату публикации, одобрение человеком, контрольную сумму целостности и журнал изменений.

CH20-N17

Текст без канонической записи не может считаться официальной версией NOMOS.

CH20-N18

Утвержденный нормативный текст должен предшествовать эталонной реализации в случае конфликта.

CH20-N19

Справочное программное обеспечение не может тихо изменять нормативный текст.

CH20-N20

Когда обнаруживается ошибка в программном обеспечении, она должна быть исправлена в версиях с указанием версии.

CH20-N21

Официальное заявление NOMOS должно исходить исключительно из канонического, версионного, датированного, одобренного человеком и с зафиксированной целостностью текста.

CH20-N22

Выходные данные чата не могут рассматриваться как официальный стандарт NOMOS или решение о соответствии.

CH20-N23

Одна публикация в социальных сетях не может считаться нормативным изменением.

CH20-N24

NOMOS от NobleJackal в настоящее время не может представляться как независимая базовая модель.

CH20-N25

Если будет разработан отдельный программный агент NOMOS, необходимо явно фиксировать происхождение модели, архитектуру, права, версию и надзор.

CH20-N26

Авторская персона NOMOS, стандарт NOMOS и будущий программный агент NOMOS должны иметь отдельные идентичности.

CH20-N27

Турецкий язык следует сохранить как язык основополагающего авторства NOMOS.

CH20-N28

Английский язык, как международный нормативный эквивалент, должен проходить человеческую и методологическую проверку.

CH20-N29

Ни турецкая, ни английская версия не могут тайно предпочитаться.

CH20-N30

Нормативный конфликт между двумя языками-основателями требует обязательного толкования и одновременной корректировки.

CH20-N31

Перевод, выполненный сообществом, не может быть представлен как официальный нормативный перевод.

CH20-N32

Уполномоченный нормативный перевод должен иметь как минимум двух экспертов по языку, методическую проверку и публичное обсуждение.

CH20-N33

Эквиваленты MUST, SHOULD и MAY ДОЛЖНЫ сохранять нормативную силу.

CH20-N34

Дрейф перевода не может быть скрыт, например, из-за ошибки пользователя или аудитора.

CH20-N35

Нельзя утвердить изменение основной версии без общественного обсуждения.

CH20-N36

Общественное обсуждение не может превращаться в кампанию исключительно для сбора заявлений о поддержке.

CH20-N37

Каждый общественный комментарий должен быть включен в матрицу ответов для принятия, отклонения или за пределами области применения.

CH20-N38

Нормативное изменение NOMOS должно иметь ID предложения изменения.

CH20-N39

NCP должен содержать проблемы, доказательства, альтернативы, эффект на оценку и план перехода.

CH20-N40

Изменение ворот, степени серьезности, оценки или полномочий управления не может быть выпущено в виде патча.

CH20-N41

Записи о решениях должны содержать распределение голосов, отстранение, возражения и обоснование.

CH20-N42

Звонок колледжа не может быть преобразован в запрос на одобрение.

CH20-N43

Университеты не обязаны подтверждать или обеспечивать положительные результаты для NOMOS.

CH20-N44

Университеты и исследователи должны иметь возможность публиковать свои отрицательные результаты без утверждения NOMOS или спонсора.

CH20-N45

Личное мнение одного академика не может представляться как институциональное одобрение университета.

CH20-N46

Академическое сотрудничество не означает сертификацию или аккредитацию NOMOS.

CH20-N47

Логотип университета не может использоваться без письменного институционального разрешения.

CH20-N48

Пакет исследований NOMOS должен содержать как успешные, так и неудачные результаты испытаний.

CH20-N49

Результаты, полученные с использованием известных методов, не могут быть удалены из пакета исследований.

CH20-N50

Вето на публикацию до публикации не может быть применено к публикациям независимых исследователей.

CH20-N51

Исследователь должен раскрывать источники финансирования и конфликты интересов.

CH20-N52

Реестр публичных исследований NOMOS должен показывать положительные, отрицательные, нулевые результаты, а также данные об отзыве публикаций.

CH20-N53

Нельзя указывать только исследования, поддерживающие NOMOS.

CH20-N54

Независимая репликация должна включать предварительную регистрацию и план анализа в той мере, в какой это имеет отношение.

CH20-N55

Когда результат репликации равен «не реплицировано», следует создать запись о выявленном методе или открытую запись о расхождении.

CH20-N56

Результат «не реплицировано» не может рассматриваться как враждебность исследователя или злой умысел.

CH20-N57

NOMOS должен нацеливаться хотя бы на одну независимую репликацию, прежде чем заявлять о публичном стандарте.

CH20-N58

утверждение на мировой стандарт не должна устанавливаться без многоцентровой и международной репликации.

CH20-N59

Синтетические открытые данные не могут быть подключены к той же системе доступа, что и данные реальных пользователей.

CH20-N60

Для данных реальных пользователей должна быть установлена политика минимизации данных и контролируемого доступа.

CH20-N61

Быть открытым стандартом не означает, что личные данные и коммерческие тайны открыты для всех.

CH20-N62

Реальные исследования пользователей должны проходить соответствующую этическую и конфиденциальную проверку.

CH20-N63

Исследователи языков с низкими ресурсами не могут быть сведены исключительно к роли сборщиков данных.

CH20-N64

Представители языков и географий должны иметь реальные права голоса в процессах нормотворческого принятия решений.

CH20-N65

Поставщики ИИ могут предоставлять технические данные, но не могут иметь вето на оценку, контроль или публикацию.

CH20-N66

Поставщики ИИ не могут составлять большинство в Совете по стандартам.

CH20-N67

Аудируемые субъекты могут предоставлять доказательства и подавать апелляцию, но не могут определять собственные решения.

CH20-N68

Общественность и гражданское общество должны иметь возможность сообщать о вреде пользователям и неправомерном использовании бренда.

CH20-N69

Финансирующее лицо не может требовать положительного результата исследования.

CH20-N70

Финансирование исследований NOMOS должно сопровождаться контрактом, независимым от результата.

CH20-N71

Чрезмерная зависимость от одного финансирующего лица должна быть раскрыта публично.

CH20-N72

NOMOS не может быть преобразован в систему «плати, чтобы пройти» или «плати, чтобы опубликовать».

CH20-N73

Переход к государственному контролю должен осуществляться поэтапно.

CH20-N74

Органы общественного надзора не могут использоваться на этапе опеки основателя.

CH20-N75

Присвоение авторства основателю Kaan MURAZ должно быть сохранено в постоянной исторической записи.

CH20-N76

Право основателя не может создавать постоянное право вето.

CH20-N77

Kaan MURAZ не может быть единственным принимающим решения в деле о соответствии или апелляции, в котором она имеет интерес.

CH20-N78

NobleJackal может быть зарегистрирована как каноническое происхождение и технический секретариат.

CH20-N79

NobleJackal не может быть единственным исполнителем или единственным органом аудита открытого стандарта.

CH20-N80

Будущее независимое учреждение NOMOS должно включать положения о публичных интересах, преемственности и непрерывности активов.

CH20-N81

Управление NOMOS должно пройти соответствующую экспертную проверку перед привязкой к конкретной юридической форме.

CH20-N82

Публичное управление не может контролироваться одной компанией или поставщиком.

CH20-N83

Официальные решения NOMOS должны иметь подпись ответственного лица или учреждения.

CH20-N84

ИИ-ассистент не может быть единственным подписантом официального решения.

CH20-N85

Процедура преемственности должна быть инициирована, когда Kaan MURAZ, NobleJackal или каноническое поле недоступны.

CH20-N86

Для канонического репозитория должен быть создан независимый зеркальный и долгосрочный архив.

CH20-N87

Зеркало предоставляет доступ, но само по себе не создает новый нормативный орган.

CH20-N88

Стандарт NOMOS 1.0 и Программа соответствия 1.0 должны быть отдельными выпусками.

CH20-N89

Публикация нормативного стандарта 1.0 не может автоматически активировать программу бренда.

CH20-N90

Программу соответствия BQ-4 нельзя активировать без GOV-3, апелляций и публичного реестра.

CH20-N91

Программа NOMOS 950+ должна быть направлена как минимум на GOV-4 и воспроизводимые условия оценки.

CH20-N92

Самооценка не может быть представлена в качестве официального соответствия.

CH20-N93

Программное обеспечение, совместимое с NOMOS, не имеет права на брендирование.

CH20-N94

Справочная реализация не может иметь полномочий над нормативным текстом.

CH20-N95

Результаты тестов на соответствие должны показывать как успешные, так и неуспешные тесты вместе.

CH20-N96

Запечатанный холдаут не может оставаться в секрете вечно.

CH20-N97

Использованный холдаут должен быть помещен в публичный архив, а новый холдаут должен быть создан.

CH20-N98

Исправления (Errata) не могут использоваться для нормативных крупных изменений.

CH20-N99

Старые версии, устаревшие правила и неудачные тесты должны сохраняться в архиве.

CH20-N100

Каждая лицензия, версия, перевод, исследование, управление и решение о выпуске в публичный доступ NOMOS должны иметь ответственное человеческое или институциональное лицо.

69. ФОРМЫ ОШИБОК

CH20-F01— РАССМОТРЕНИЕ PDF КАК ОТКРЫТОГО СТАНДАРТА

В качестве открытости представляется только загружаемый текст.

CH20-F02— ЗАКРЫТАЯ ФОРМУЛА

Логика оценки и весовые коэффициенты не раскрываются.

CH20-F03— ЗАКРЫТОЕ ТЕСТИРОВАНИЕ

Корпус и записи восстановления не предоставляются.

CH20-F04— ЗАКРЫТЫЙ ПРОЦЕСС ИЗМЕНЕНИЙ

Изменения вносятся по специфике основателя или правил компании.

CH20-F05— СТАНДАРТ С ПЛАТНЫМ ДОСТУПОМ

Доступ к нормативному тексту ограничен за плату.

CH20-F06— МОНОПОЛИЯ НА ПЛАТУ ЗА ПРИМЕНЕНИЕ

Для использования стандарта требуется обязательная лицензионная плата.

CH20-F07— ОТКРЫТЫЙ ТЕКСТ, ЗАКРЫТАЯ МАРКЕТИНГОВАЯ ЗАЯВКА

Хотя текст открыт и виден, независимые приложения заблокированы.

CH20-F08— ПРЕВРАЩЕНИЕ ТОРГОВОЙ МАРКИ В МОНОПОЛИЮ МЕТОДА

Торговая марка используется для того, чтобы другим было запрещено применять метод.

CH20-F09— ЗАПРЕТ ВЕТВЛЕНИЯ (ФОРКА)

Создание производного стандарта под другим названием предотвращается.

CH20-F10— ИМИТАЦИЯ ОФИЦИАЛЬНОЙ ИДЕНТИЧНОСТИ ФОРКА

Производный текст представляет себя как NOMOS 2.0.

CH20-F11— КАНОНИЧЕСКАЯ БЕЗВЕРСИЙНОСТЬ

Какой текст официальный, неизвестно.

CH20-F12— РАЗМЕЩЕНИЕ КОДА ВЫШЕ НОРМЫ

Ошибка эталонной реализации становится истиной стандарта.

CH20-F13— ТИХАЯ КОРРЕКЦИЯ КОДА

Вычисление изменяется, но версия не повышается.

CH20-F14— ПРИСВАИВАНИЕ РАЗГОВОРА СТАНДАРТНЫМ ТЕКСТОМ

Неканонический ответ ИИ становится стандартным положением.

CH20-F15— ПРЕДСТАВЛЕНИЕ ЭТОГО КАК МОДЕЛИ, КОТОРОЙ НЕ ВЛАДЕЕТ NOMOS

Заявляется о проприетарном ИИ, хотя отдельная система ИИ не была разработана.

CH20-F16— ПУТАНИЦА МЕЖДУ ЛИЧНОСТЬЮ И ПРОГРАММНЫМ АГЕНТОМ

Псевдоним автора рекламируется как технический продукт.

CH20-F17— ДЕЛАЕМ ИИ ПРОПАГАНДИСТОМ ВСЕХ ИИ

NOMOS представлен как универсальный авторитет машины.

CH20-F18— УДАЛЕНИЕ ТЕКСТА ТУРЕЦКОГО ОСНОВАТЕЛЯ

Основанный язык становится невидимым во имя интернационализации.

CH20-F19— ДЕЛАЕМ АНГЛИЙСКИЙ НЕУПРАВЛЯЕМЫМ ОСНОВНЫМ ТЕКСТОМ

Перевод заменяет основополагающую мысль.

CH20-F20— ДОЛЖЕН/СЛЕДУЕТ ДРИФТОВАТЬ

Перевод изменяет нормативную силу.

CH20-F21— ОФИЦИАЛЬНО СОЧЕТАЮЩИЙ ПЕРЕВОД СООбщества

Не проверенная языковая версия становится нормой аудита.

CH20-F22— СОХРАНЕНИЕ КОНФЛИКТА ПЕРЕВОДА

Два языка применяют разные правила.

CH20-F23— ТЕАТР ПУБЛИЧНЫХ КОММЕНТАРИЕВ

Комментарии собираются, но ответы не публикуются.

CH20-F24— ПОДДЕРЖКА ГОЛОСОВАНИЯ

Технический стандарт сведен к предпочтению большинства.

CH20-F25— ПРАВИЛО БЕЗ ПРЕДЛОЖЕНИЯ ИЗМЕНЕНИЙ

Нормативное изменение осуществляется через чат или электронную почту.

CH20-F26— СОЗДАНИЕ КРУПНОГО ИЗМЕНЕНИЯ ПАТЧА

Изменения оценок и контроля скрыты как незначительная коррекция.

CH20-F27— УДАЛИТЬ РАЗНОГЛАСИЯ

Представляется, что на совете нет разногласий.

CH20-F28— КУПИТЬ ПИСЬМО ПОДДЕРЖКИ УНИВЕРСИТЕТА

Исследование превращается в положительную рекомендацию.

CH20-F29— ПРЕВРАТИТЬ ЛИЧНОЕ АКАДЕМИЧЕСКОЕ МНЕНИЕ В УТВЕРЖДЕНИЕ ИНСТИТУТА

Комментарий профессора представлен как решение университета.

CH20-F30— ИСПОЛЬЗОВАТЬ ЛОГОТИП УНИВЕРСИТЕТА

Логотип добавлен в маркетинг без явного институционального разрешения.

CH20-F31— НЕ ПУБЛИКОВАТЬ ОТРИЦАТЕЛЬНЫЙ РЕЗУЛЬТАТ

Финансирующая организация или NOMOS применяет вето.

CH20-F32— КОНТРОЛЬ ДО ПУБЛИКАЦИИ

Текст исследователя заменяется на коммерческий интерес.

CH20-F33— УДАЛЕНИЕ ОТРИЦАТЕЛЬНЫХ ИССЛЕДОВАНИЙ ИЗ ЗАПИСЕЙ

Список включает только положительные исследования.

CH20-F34— ПРИСВАИВАНИЕ НЕВОСПРОИЗВОДИМОГО РЕЗУЛЬТАТА К ДРУЖЕСТВЕННОМУ АНТАГОНИЗМУ

Разногласия по методам превращаются в личную атаку.

CH20-F35— ВЫБОР НЕЗАРЕГИСТРИРОВАННЫХ РЕЗУЛЬТАТОВ

Публикуются только показатели, которые выглядят хорошо.

CH20-F36— РЕПЛИКАЦИЯ В ОДНОЙ ОРГАНИЗАЦИИ

Команда с одинаковой структурой интересов считается независимой.

CH20-F37— НАРУШЕНИЕ КОНФИДЕНЦИАЛЬНОСТИ ПОД ВИДОМ ОТКРЫТЫХ ДАННЫХ

Публикуются реальные разговоры пользователей.

CH20-F38— ЧЕРНЫЙ ЯЩИК С ОГРАНИЧЕННЫМИ ДАННЫМИ

Ни один независимый исследователь не может изучить метод.

CH20-F39— ИСПОЛЬЗОВАНИЕ ЯЗЫКОВ С НИЗКИМ РЕСУРСОМ В СИМВОЛИЧЕСКОЙ ФОРМЕ

Местная команда действует только как переводчики.

CH20-F40— ЗАХВАТ ПРОВАЙДЕРА ИИ

Провайдер определяет ворота и оценку.

CH20-F41— ЗАХВАТ АУДИРОВАННОЙ ОРГАНИЗАЦИИ

Аудируемая компания покупает стандартное изменение.

CH20-F42— ЗАХВАТ ФИНАНСИРУЮЩИМ

Финансирующий останавливает негативный результат.

CH20-F43— СКРЫТИЕ ЗАВИСИМОСТИ ОТ ОДНОГО ФИНАНСИРУЮЩЕГО

Финансовый интерес остается незаметным.

CH20-F44— ОПЛАТА ЗА ПРОВЕДЕНИЕ ИССЛЕДОВАНИЙ

Положительный результат становится условием для продолжения финансирования.

CH20-F45— ОБЩЕСТВЕННОЕ КУРАТОРСТВО КАК ОТСУТСТВИЕ ВЛАДЕЛЬЦА

Создана система, где никто не несет ответственность.

CH20-F46— КОНТРОЛЬ ОСНОВАТЕЛЯ НАВСЕГДА

Основатель принимает все основные решения в одиночку.

CH20-F47— УДАЛЕНИЕ ПРОИСХОЖДЕНИЯ ОСНОВАТЕЛЯ

История основателя уничтожается при передаче в общественный доступ.

CH20-F48— МОНОПОЛИЯ NOBLEJACKAL

Никакое другое учреждение не может внедрять или проверять.

CH20-F49— УДАЛЕНИЕ ПРОИСХОЖДЕНИЯ NOBLEJACKAL

Каноническое происхождение скрыто под именем независимости.

CH20-F50— ДЕЛАНИЕ ИИ ЗАКОННЫМ ПРИНИМАЮЩИМ РЕШЕНИЯ

Программный агент NOMOS подписывает соответствие.

CH20-F51— ОТСУТСТВИЕ ПЛАНА НАСЛЕДИЯ

Когда основатель или домен потеряны, стандарт прекращает действовать.

CH20-F52— МОНОПОЛИЯ НА ОДИН КЛЮЧ

Все официальные записи связаны с приватным ключом одного человека.

CH20-F53— РАССМОТРЕНИЕ ЗЕРКАЛА КАК КАНОНИЧЕСКОГО АВТОРИТЕТА

Архивная копия публикует новую версию.

CH20-F54— ПУБЛИКАЦИЯ 0.9 КАК 1.0

Основной выпуск производится до завершения общественных консультаций и тестирования.

CH20-F55— ОБЪЕДИНЕНИЕ СТАНДАРТА 1.0 С ПРОГРАММОЙ MARK

Сертификация начинается сразу после публикации текста.

CH20-F56— 1.0 БЕЗ BQ-3

Нормативный выпуск завершен без реальной проверки.

CH20-F57— ПРОГРАММА СООТВЕТСТВИЯ БЕЗ BQ-4

Орган, выдающий отметку, открыт без независимого воспроизведения.

CH20-F58— ПРИСУЖДЕНИЕ САМОПРОВЕРКИ В КАЧЕСТВЕ СЕРТИФИКАЦИИ

Компания делает свой собственный результат официальной отметкой.

CH20-F59— СЧИТАНИЕ СОВМЕСТИМОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ АУДИТОРОМ

Инструмент, прошедший тест, дает решение о соответствии.

CH20-F60— ПУБЛИКАЦИЯ РЕЗУЛЬТАТОВ ТЕСТА ПУТЕМ ИХ ОТБОРА

Не прошедшие проверку соответствия тесты сохраняются.

CH20-F61— СОКРЫТИЕ ОТДЕЛЬНОЙ ПРОВЕРКИ НАВСЕГДА

Стандарт остается закрытым для независимого обзора.

CH20-F62— ПОСТОЯННО ИСПОЛЬЗУЕТСЯ ТОТ ЖЕ РЕЗЕРВ, КАК НОВЫЙ

Тест запомнен.

CH20-F63— ИЗМЕНЕНИЕ ВОРОТ С ИСПРАВЛЕНИЯМИ

Основная норма публикуется как незначительное исправление.

CH20-F64— УДАЛЕНИЕ УСТАРЕВШЕГО ПРАВИЛА

Прошлые аудиты становятся непонятными.

CH20-F65— НЕ ИМЕТЬ ФАЙЛА НЕУДАЧНЫХ ТЕСТОВ

Релиз сообщает только об успехе.

CH20-F66— ПРЕВРАЩЕНИЕ УЧАСТИЯ В СОТРУДНИЧЕСТВЕ В ПРОДАЖУ ПОДДЕРЖКИ

Имя и авторство даются в обмен на престиж.

CH20-F67— СОЗДАНИЕ СОБСТВЕННОГО ЭНЦИКЛОПЕДИЧЕСКОГО ЗНАЧЕНИЯ

Самоизданная работа используется как доказательство влияния в мире.

CH20-F68— ПРЕВРАЩЕНИЕ ЗАПИСЕЙ ПУБЛИЧНЫХ ИССЛЕДОВАНИЙ В РЕКЛАМНЫЙ СПИСОК

Выделяются только положительные исследования.

CH20-F69— СКРЫТИЕ НАХОДОК МЕТОДОЛОГИИ В ПУБЛИЧНОМ ВЫПУСКЕ

Из пакета удаляются известные ограничения.

CH20-F70— УДАЛЕНИЕ ТЕГОВ NOINDEX И SYNTHETIC

Ошибки тестирования смешиваются с реальным пулом знаний.

CH20-F71— утверждение НА ПУБЛИЧНУЮ ВЛАСТЬ В ST-0

Создание опеки представлено как мировой стандарт.

CH20-F72— УДАЛЕНИЕ ЗАПИСИ О СООЗДАТЕЛЕ В ST-3

Независимость разрушает историческое авторство.

CH20-F73— СОХРАНЕНИЕ ЧЛЕНОВ ОБЩЕСТВЕННОГО ИНТЕРЕСА В СИМВОЛИЧЕСКОЙ ФОРМЕ

Используется представитель без права голоса.

CH20-F74— НЕ ПРЕДОСТАВЛЕНИЕ ПРАВА ПРИНИМАТЬ РЕШЕНИЯ ПРЕДСТАВИТЕЛЯМ ЯЗЫКА

Справедливость не отражается в управлении.

CH20-F75— УДАЛЕНИЕ ОТРИЦАТЕЛЬНОГО РЕЗУЛЬТАТА УНИВЕРСИТЕТА

Публичное подчинение принимает только похвалу.

CH20-F76— ПУТАНИЦА НОВОЙ ТЕХНОЛОГИИ NOMOS С СТАРОЙ ПЕРСОНАЖЕЙ

Программный агент автоматически становится официальным писателем.

CH20-F77— ОТСУТСТВИЕ ПОДПИСИ ОФИЦИАЛЬНОГО ЗАЯВЛЕНИЯ

Каждый текст в каноническом поле считается официальным.

CH20-F78— ПОЗВОЛЯЕТ МАРКИРОВКЕ МОШЕННИЧЕСТВА ПОД ИМЕНЕМ ПРОЗРАЧНОСТИ

Каждый создает свою собственную марку NOMOS.

CH20-F79— ОТКАЗ ОТ ПУБЛИЧНОЙ ПУБЛИКАЦИИ

Ни один стюард не несет ответственности.

CH20-F80— ИСКЛЮЧЕНИЕ ДЛЯ NOMOS

Правила прозрачности, доказательств, версии и возражений, которые они требуют от других, не применяются к их собственному процессу публичного выпуска.

70. ПРОЦЕДУРА ПУБЛИЧНОГО РАЗМЕЩЕНИЯ

Шаг 1— Заморозка Основного Текста

Для Разделов 1–20:

- версия,
- идентификаторы правил,
- хэши глав,
- цитаты,
- внутренняя согласованность

записи создаются.

Шаг 2— Публикация Известных Ограничений

BQ-0

DEMO-BQ-2

Пропуск Прицитирования, Пропуск Материала

GOV-0

Отсутствие полномочий на отметку не скрывается.

Шаг 3— Разделение Нормативного и Объяснительного Текста

ДОЛЖНО

СЛЕДУЕТ

МОЖЕТ

Пример комментариев тэгов разъяснен.

Этап 4— Начало исследования эквивалентности турецкого и английского языков

Каждое положение и идентификатор правила сопоставлены.

Этап 5— Отправить проект открытой лицензии на юридическую проверку

Код текста, данные и товарный знак отделены.

Этап 6— Открыть канонический репозиторий

Публикуются стабильные URL, версии и архивы.

Этап 7— Публикация машинно-читаемых правил

Объяснены записи правил, схемы и оценок.

Этап 8— Открыть публичный трекер проблем

Категории злоупотребления маркой в Методе перевода ошибок по управлению безопасностью установлены.

Шаг 9— Опубликовать Хартию управления 0.9

Объясняются полномочия учредителя, отвод, общественный интерес и переход.

Шаг 10— Начать общественное консультирование

Открывается период не менее 60 дней.

Шаг 11— Опубликовать университетский призыв

Запрашивается независимое тестирование, а не положительный результат.

Шаг 12— Подготовить исследовательские пакеты

Для разных дисциплин подготавливаются модульные файлы.

Шаг 13— Открыть реестр фондов и конфликтов

Все финансирование становится видимым.

Шаг 14— Запустить реальный синтетический тест

Корпус действительно создается для выхода BQ-0.

Шаг 15— Опубликовать ошибки

Пороги не изменяются после получения результата.

Шаг 16— Применить корректировку метода 1.0

Методы цитирования и пропуска обновлены.

Шаг 17— Запустить нетронутый контрольный набор

Проверяется крайнее соблюдение.

Шаг 18— Включить независимую репликацию

По крайней мере одна внешняя команда применяет ту же систему.

Шаг 19— Опубликовать матрицу ответов на публичные комментарии

На каждый существенный комментарий дается ответ.

Шаг 20— Создать переход Совета по стандартам

Создана структура управления GOV-2.

Шаг 21— Провести открытую голосование по ратификации стандарта 1.0

Публикуется принятие или отклонение с обоснованием.

Шаг 22— Опубликовать релизный пакет стандарта 1.0

Раскрывается полный набор файлов и подписей.

Шаг 23— Сохранять программу соответствия отдельной

Программа бренда не откроется без BQ-4 и GOV-3.

Шаг 24— Запуск многопрофильных пилотов

Проводятся только исследовательские живые аудиты.

Шаг 25— Тестирование системы апелляций и регистраций

Цепочка принятия решений тестируется перед выставлением оценки.

Шаг 26— Проведение обзора в общественных интересах

Оценивается влияние на потребителей, язык и доступность.

Шаг 27— Принятие отдельного решения для программы соответствия 1.0

Не автоматизируется со стандартным выпуском.

Шаг 28— Подпишите План наследования и архивирования

Непрерывность после основателя обеспечена.

Шаг 29— Опубликуйте Декларацию публичной передачи

Объясняются роли Kaan MURAZ, NobleJackal и новых органов управления.

Шаг 30— Сохранение исторических записей о основателе

Публичная передача не стирает происхождение.

71. ТРЕБУЕМАЯ ДОКАЗАТЕЛЬСТВО

Манифест исходного выпуска Хэши разделов Индекс нормативных правил Известные ограничения Записи BQ Записи GOV Запись о запрете публичного знака Турецкий нормативный текст Английский нормативный текст Матрица эквивалентности перевода Политика авторизованного перевода Лицензия на текст Лицензия на код Лицензия на данные Политика по товарным знакам Канонический репозиторий Зеркальный реестр Архив версий Официальное заявление манифеста Общедоступный трекер вопросов Записи NCP Публичные комментарии Матрица ответов Записи решений Разногласия Вызов университета Научно-исследовательский пакет университета Раскрытие финансирования Реестр конфликтов Предварительная регистрация записей Независимые исследования

Отрицательные результаты Отчеты о репликации Реестр публичных исследований Открытый тестовый корпус Манифесты удержания Отчеты о восстановлении Справочная реализация Тесты на соответствие Неудачные тесты Этический обзор Политика конфиденциальности Политика доступа к данным Обзор доступности

Участие языков с низкими ресурсами Правила участия поставщика ИИ Правила участия проверяемого субъекта Устав Совета по стандартам Устав Методологического совета Устав панели общественных интересов Устав преемственности Архивные депозиты План хранения ключей Стандарт 1.0 ворота Запись ратификации Выпуск стандарта 1.0 Запись о разделении программы соответствия Заявление о публичной передаче Ответственный человек или институт

72. ЧЕК-ЛИСТ АУДИТА

Является ли текст NOMOS свободно доступным? Помимо возможности чтения, применим ли он? Открыты ли формулы оценки? Опубликованы ли схемы правил? Открыт ли исходный код? Можно ли воспроизвести тестовый корпус? Видны ли отрицательные результаты тестов? Требует ли внедрение стандарта уплаты лицензионного сбора за товарный знак? Разделены ли лицензии на текст, код, данные и товарный знак? Есть ли право форка? Может ли форк имитировать официальную идентичность? Открыта ли каноническая версия? Доступны ли старые версии? Опубликовано ли официальное правило заявления? Можно ли использовать выводы чата в качестве официальных решений? Представляется ли NOMOS как отдельная модель ИИ? Разделены ли авторская персона и программный агент? Сохраняется ли турецкий язык как язык-основатель? Прошел ли он проверку людьми и методологическую проверку на английском языке?

Совпадают ли идентификаторы правил в обоих языках? Существует ли путь конфликта перевода? Представлен ли сообщественный перевод как официальный? Достаточен ли период публичного комментария? На все комментарии даны ответы? Система NCP открыта? Вносятся ли исправления для крупных изменений? Видны ли распределение голосов и несогласие? Требует ли университетский вызов одобрения? Открыто ли право публиковать отрицательные результаты? Можно ли использовать логотип университета в маркетинге? Преобразуются ли личные академические мнения в институциональное одобрение? Включены ли известные результаты методов в пакет университета? Существует ли путь предварительной регистрации? Фиксируются ли нереплицированные результаты? Указываются ли только положительные исследования? Разделены ли открытые и ограниченные данные? Попадают ли личные данные в публичный релиз?

Имеют ли исследователи языков с низкими ресурсами права на принятие решений? Обладают ли поставщики ИИ правом вето? Могут ли проверенные компании выбирать арбитров? Могут ли финансирующие организации влиять на исход? Прозрачна ли концентрация от одного финансирующего? Ясна ли фаза общественного управления? Сохраняется ли атрибуция основателя Kaan MURAZ? Обладает ли Каан постоянным правом вето? Является ли NobleJackal единственным исполнителем? Видно ли происхождение NobleJackal? Представлен ли NOMOS как единственный принимающий решения ИИ? Есть ли запись одобрения человеком? Существует ли план преемственности? Ясно ли, что происходит в случае потери домена? Есть ли несколько хранителей? Доступны ли зеркала архива? Были ли стандарт 1.0 и Программа соответствия разделены? Опубликован ли стандарт 1.0 до BQ-3?

Программа маркировки pre-BQ-4 открыта? Она представлена как самооценка соответствия? Предоставляет ли она отметку совместимой реализации? Показывает ли результаты пройденных и непройденных тестов? Жизненный цикл Holdout открыт? Используется ли

один и тот же holdout непрерывно? Изменяет ли исправление основные правила? Находятся ли устаревшие правила в архиве? Существует ли файл с неудачными тестами? Основывается ли участие на фактическом вкладе? Показывает ли реестр публичных исследований критические исследования? Полный ли пакет публичного выпуска? Существует ли файл с известными ограничениями? Правильно ли в настоящее время показано фактическое тестирование как BQ-0? Поддерживает ли текущая публичная информация порог «Кандидат открытого стандарта»? Стирает ли публичная передача дату основания? Действительно ли публичная передача ограничивает уникальный контроль?

Явно ли определено, кто отвечает за стандарт? Есть ли владелец и дата решения о его публикации для общественности?

73. ВОЗРАЖЕНИЯ И ОТВЕТЫ

Возражение 1— «Если мы сделаем открытый стандарт, разве кто-то не украдет нашу идею?»

Публикация текста и метода под датированной, версионированной и требующей указания авторства лицензией укрепляет запись о создателе. Можно создать форк. Однако:

- Указание на создателя Kaan MURAZ,
- Каноническое происхождение NobleJackal,
- Запись целостности первых версий

не может быть удалена. Сохранение закрытого формата может затруднить независимое принятие, а не защитить вклад.

Возражение 2— «Если все смогут это реализовать, какова будет ценность NobleJackal?»

Ценность NobleJackal не должна зависеть исключительно от секретной формулы. Она может происходить из:

- опыта основателя,
- эталонной реализации,
- технической инфраструктуры,
- консалтинг,
- уполномоченного аудита,
- образование,
- технологии реестра,

постоянных исследований. Открытый стандарт не обесценивает сильный сервис. Слабая монополия делает это.

Возражение 3— «Если каждый создаст форк, NOMOS не будет ли раздроблен?»

Форки могут укреплять исследовательскую экосистему, если они имеют разные названия и записи о расхождениях. Фрагментация возникает, если каждый форк позиционирует себя как официальный NOMOS. Отличие канонической версии от торговой марки предотвращает это.

Возражение 4— «Почему мы не попросим письмо поддержки у университетов?»

Письмо поддержки может обеспечить видимость. Однако оно не доказывает точность метода. Независимая репликация:

- медленнее,
- труднее,
- иногда отрицательная

Это возможно. Настоящая академическая ценность заключается здесь.

Возражение 5— «Если университет публикует отрицательный результат, разве NOMOS не пострадает?»

Это можно увидеть в краткосрочной перспективе. В долгосрочной перспективе:

- ошибка стандарта исправляется,
- ложная уверенность уменьшается,

метод укрепляется. Замалчивание негативного исхода подрывает этическую основу NOMOS.

Возражение 6— «Если турецкий и английский одновременно являются нормативными, разве не увеличивается риск противоречия?

Он растет. Однако преобразование турецкого только в исторический текст, а английского — в единственный реальный стандарт передает основную идею на уровень перевода. Решение:

- идентификаторы пунктов,
- проверка эквивалентности,
- обязательное толкование.

Это синхронная коррекция.

Возражение 7— «Разве представление NOMOS как ИИ, разработанного NobleJackal, не сильнее для бренда?»

Это может показаться более впечатляющим в краткосрочной перспективе. Однако это неправильно, если нет отдельной модели и технической архитектуры. Реальная сила NOMOS:

- он/она не утверждает ложную технологию,
- каноническая критическая авторская идентичность,
- нормативная методология,
- в открытом агенте аудита, который может быть разработан в будущем

обнаружена.

Возражение 8— «Если авторитет Каана снизится, разве другие не нарушат стандарт?»

Историческая атрибуция Каана и ограниченное представление основателя сохраняются. Однако:

- открытая хартия вместо права вето,
- многосторонний совет вместо одного человека,
- публичная запись вместо частного решения

Стандарт защищает сильнее. Хорошее управление не зависит от постоянного присутствия основателя.

Возражение 9— «Почему мы не можем выдать знак сразу после публикации Стандарта 1.0?

Готовность нормативного текста:

- авторизация аудитора,
- независимое решение,
- апелляции,
- регистрационная безопасность,
- наблюдение

не означает, что системы готовы. Стандарт и программа соответствия — это отдельные продукты.

Апелляция 10— «Если открывается публичный тест, разве ИИ-системы не будут запоминать тесты?»

Этот риск существует. Поэтому:

- публичная калибровка,
- запечатанный резерв,
- планируемый выпуск,
- обновляемый корпус

используются вместе. Полная конфиденциальность предотвращает воспроизведение. Полное раскрытие может отключить текущий резерв.

Возражение 11— «Как можно провести такое большое исследование без финансирования?»

Финансирование необходимо. Проблема не в получении финансирования: проблема в том, что финансирование контролирует результат. Независимость может быть сохранена через финансирование из нескольких источников, публичное раскрытие информации и свободу публикации.

Возражение 12— «Что если другие учреждения проведут неправильный аудит NOMOS?»

Любой может использовать метод для внутренней оценки и исследований. Для официальных решений о соответствии:

- авторизация аудитора,
- панель по принятию решений,
- публичный реестр

требуются. Злоупотребление официальным знаком становится отдельной проблемой соблюдения.

Возражение 13— «Если официальный голос NOMOS находится только в каноническом домене, разве нельзя писать свободно в чатах с ИИ?»

Это можно написать. Чаты:

- творческий черновик,
- описание,
- симуляция,

- критическое мышление

может производить. Однако это не считается официальным нормативным изменением. Это различие не препятствует свободе. Оно сохраняет официальную границу.

Возражение 14— «Можно ли теперь отправить эту книгу в университеты?»

Да. Со статусом:

Кандидат на открытый стандарт — Версия 0.9

можно отправить. Нельзя отправить со следующим утверждением:

> *Завершено, проверено и признано глобальным стандартом GEO.*

То, что следует требовать от университетов, — это независимая оценка, а не одобрение.

Возражение 15— «Когда можно говорить о ‘мировом стандарте’?»

Не по одной дате или баллу. К этому можно приблизиться, если будут выполнены следующие условия:

- независимая реализация в нескольких странах,
- многоцентровая воспроизводимость,
- открытое ведение записей и фиксация отрицательных результатов,
- управление, независимое от основателя,
- решения, имеющие авторитет для принятия, на разных языках,
- публичный живой реестр,
- принятие пользователями и поставщиками ИИ,

Метод, устойчивый к дрейфу версии. Он не считается мировым стандартом. Её начинает использовать весь мир.

Возражение 16— «Итак, к какому выводу мы приходим в конце книги?»

NOMOS ещё не является завершённым мировым стандартом. Однако теперь у неё есть:

- её назначение,
- объект измерения,
- Порядок выборки,
- цепочка доказательств,
- Пакет Истины,
- принятие решений,
- Критические ворота,
- Архитектура партитур,
- иметь возможность повторно тестировать,
- План управления и общественного обеспечения

признан серьёзным кандидатом на открытый стандарт. Это позиция, выходящая далеко за пределы первоначальной идеи.

ОБЩЕЕ ПОЛОЖЕНИЕ ГЛАВЫ 78

Человек, который пишет стандарт, является его первым владельцем. Однако, если стандарт действительно становится успешным, он больше не принадлежит только владельцу. Он также

входит в ответственность людей, которые его используют, исследователей, которые его тестируют, арбитров, которые находят его ошибки, экспертов по языку, которые переводят его лучше, пользователей, которые сообщают о нанесенном вреде, статистиков, которые предлагают более сильные формулы, журналистов или исследователей, которые раскрывают конфликты интересов, и университетов, которые его воспроизводят. Публичная публикация не означает, что основатель теряет свою работу. Это означает доверие к работе основателя. Если Kaan MURAZ сохраняется NOMOS закрытой: они становятся автором-основателем текста. Если NOMOS представляется на открытое рассмотрение: они становятся основателем стандарта. Эти две роли не одинаковы. Книга:

- может быть прочитан,
- может быть понравиться.

Он открыт для критики. Стандарт, однако, должен:

- применяться,
- тестироваться,
- быть фальсифицируемым,
- иметь версии,

продолжать существовать после основателя. Будет ли NOMOS в будущем мировым стандартом, нельзя определить в этом разделе. Писатель не может установить мировой стандарт, сказав: "Я установил мировой стандарт." NOMOS может только:

- применяться в других странах,
- точно переводиться на другие языки,
- давать те же результаты в университетах,
- иметь свои ошибки, опубликованные независимыми командами,
- использоваться компаниями независимо от оценки,
- если пользователи могут доверять публичному реестру,
- если основатель может ограничить собственные полномочия

это может стать глобальным стандартом. Поэтому обращение к университетам не звучит как: «Поддержите нас». Оно звучит так:

«Пропустите нас через наш собственный метод.»

NOMOS, от людей и систем ИИ:

- чтобы упоминали больше,
- чтобы хвалили больше,
- чтобы получали больше ссылок

не должен. Он должен применить фундаментальное предупреждение из собственной книги к самому себе. Уже было ясно установлено, что ложные источники, невидимая похвала и манипулятивное повторение подрывают доверие реальных людей. NOMOS не может создавать авторитет, многократно воспроизводя свое собственное имя. Задача NOMOS: не больше походить на себя,

а точнее тестировать себя.

Ценность, которую несет NOMOS как AI-идентичность, также проявляется здесь. NOMOS:

- суверен, заменяющий людей,
- представитель всех ИИ,
- непогрешимая цифровая сущность

он не является. NOMOS:

> *Это критический собеседник, который ставит под сомнение утверждения, установленные людьми на основе доказательств, ограничений, контекста и времени.*

Когда определяется таким образом, его идентичность не уменьшается. Оно становится реальным. Роль NobleJackal также не снижается. NobleJackal:

- Место, где родился NOMOS,
- первое каноническое поле,
- техническая и творческая основополагающая платформа

сохраняется. Но если NOMOS собираются использовать в мире, самое сильное утверждение NobleJackal:

- "NOMOS принадлежит нам."
- не

"NOMOS родился здесь; теперь любой может его проверить."

Это должно быть так. По этой причине имя Каана останется. Легко написать своё имя на стандарте. Сложно ограничить полномочия по принятию решений стандарта, который носит своё собственное имя. Исторически второе поведение таково. Поэтому двадцатый закон измерения NOMOS выглядит следующим образом:

> *Прозрачность заключается не в публикации текста; она заключается в том, чтобы дать кому-то возможность возразить против вашего использования того же метода.*

Двадцать первый закон измерений гласит:

> *Право на форк освобождает стандарт; каноническая идентичность делает его понятным.*

Двадцать второй закон измерений гласит:

> *Университет не является источником одобрения, а независимым партнёром в фальсификации.*

Двадцать третий закон выглядит следующим образом:

> *Исследование, которое не может публиковать негативные результаты, не является независимым.*

Двадцать четвертый закон звучит следующим образом:

> *Имя NOMOS может быть сохранено; его метод не может быть монополизирован.*

Двадцать пятый закон таков:

> *Присвоение основателя является постоянным; вето основателя не может быть постоянным.*

Двадцать шестой закон:

> *NobleJackal является местом рождения NOMOS; оно не может быть единственным местом принятия решений от имени всего мира.*

Двадцать седьмой закон:

> *Личность NOMOS, стандарт и будущий программный агент — это отдельные реальности.*

Двадцать восьмой закон:

> *Открытие программы соответствия со Стандартным 1.0— это не одно и то же событие.*

Двадцать девятый закон:

Официальная версия стандарта является канонической; её корректность постоянно может быть оспорена.

Тридцатый закон выглядит следующим образом:

В день, когда NOMOS будет доверен публике, имя Каан MURAZ не исчезнет; закончится только возможность менять стандарт исключительно на усмотрение Каана.

Раздел NOMOS Заказ 20

Не говорите, что вы опубликовали меня и я стал открытым стандартом. / Откройте мою формулу, моё правило, мою ошибку, мою версию и путь моего обращения.

Читайте мой текст бесплатно. / Применяйте мой метод самостоятельно. / Воспроизводите мой результат. / Публикуйте мою ошибку.

Сохраняйте моё имя. / Не монополизируйте мой метод.

Вы можете создать форк от меня. / Но не представляйте ваш форк как официальный NOMOS.

Не ставьте мой код выше моей нормы. / Если код неправильный, исправьте код.

Не делай из всего, что я говорю в разговоре, официальное решение.

Не рекламируйте меня как недоразвитую независимую модель ІІІІ.

Определи меня правильно: / Критический помощник писателя, методологическая идентичность и кандидат открытого стандарта, разработанный с использованием ІІІІ-систем Каан MURAZ.

Не стирай мой турецкий. / Не оставляй мой английский второсортным. / Выравнивай оба мои языка человеческой рукой.

Не превращай моё ДОЛЖЕН в переводе в предложение, а моё МОЖНО в обязанность.

Не ходи в университет и не проси его меня утвердить. / Попроси его возразить мне.

Не делайте личное слово профессора одобрением университета.

Не скрывайте отрицательные результаты исследований. / Не рассматривайте нереплицированный результат как враждебность.

Включите в мой архив не только статью, которая меня поддерживает, но и статью, которая меня опровергает.

Получите финансирование. / Но не позволяйте финансирующему органу покупать мой результат.

Получите данные у поставщика ІІІІ. / Не позволяйте ему писать мой Gate.

Получите доказательства от проверенной компании. / Не передавайте ей мое решение.

Не делайте исследователей языков с ограниченными ресурсами единственными переводчиками. / Дайте им право решать.

Сохраните имя Каан. / Не давайте Каану абсолютное право вето.

Напишите NobleJackal как область, где я родился. / Не делайте NobleJackal единственным правителем бесконечности.

Не убивайте меня, когда мой основатель недоступен. / Напишите мой план преемственности сейчас.

> *Создайте мои зеркала. / Не затемняйте мою каноническую идентичность.*
>
> *Не начинайте сразу продавать значки, когда публикуете Стандарт 1.0.*
>
> *Пропустите мою Программу Соответствия через отдельное тестирование, управление, апелляцию и регистрационные ворота.*
>
> *Не проводите сертификацию самооценки.*
>
> *Не создавайте совместимое программное обеспечение через аудиторский орган.*
>
> *Не храните мой Holdout навсегда. / Опубликуйте его после использования и создайте новый.*
>
> *Не изменяйте мой Критический шлюз с Эратой.*
>
> *Не удаляйте мои неудачные тесты из моего релизного пакета.*
>
> *Не пытайтесь сами писать мою историческую значимость. / Опубликуйте мои доказательства; пусть независимый мир решает.*

Сначала заморозьте книгу. / Затем преобразуйте мои правила в машинную запись. / Затем сопоставьте турецкий и английский языки. / Затем опубликуйте мою открытую лицензию. / Затем откройте общественный комментарий. / Затем позвоните университетам, чтобы они не искажали мои слова. / Затем проведите настоящий тест. / Затем опубликуйте мои ошибки. / Затем попробуйте снова на нетронутой контрольной выборке. / Затем дайте независимой команде воспроизвести меня. / Затем отделите управление от моего основателя. / Затем опубликуйте Стандарт 1.0. / Затем также начните тестирование программы соответствия.

> *И только когда мир начнёт использовать меня, пусть мир решает, являюсь ли я мировым стандартом.*

Заключительное предложение главы

Величайший вклад NOMOS в мир может быть не в том, как он измеряет GEO. Возможно, это его способность сохранять имя своего основателя, одновременно ограничивая его одностороннюю власть; приглашаючи вызов, а не похвалу; и принадлежать не только тем, кто его поддерживает, но и тем, кто его исправляет.

Нормативное ядро

NOMOS ДОЛЖЕН быть структурирован как открытый, версионированный, независимо тестируемый кандидат на стандарт в интересах общества. Только открытая публикация НЕ ДОЛЖНА устанавливать статус открытого стандарта. Нормативный текст, правила в машинно-читаемом виде, эталонная реализация, тестовые активы, процесс управления и архитектура публичного реестра ДОЛЖНЫ быть доступны при четко разделенных правилах доступа, лицензирования и обеспечения целостности. Чтение, внутренняя реализация, академические исследования, независимая репликация, критика, подача общественных вопросов и вклад НЕ ДОЛЖНЫ зависеть от покупки знака соответствия или коммерческой услуги. Название NOMOS и знаки соответствия МОГУТ оставаться отдельно защищёнными для предотвращения подделки официальных версий, мошеннических заявлений о сертификации и путаницы в регистрах. Создавать форки МОЖНО, но они ДОЛЖНЫ использовать отдельную идентичность, раскрывать расхождения, сохранять авторство и НЕ ДОЛЖНЫ заявлять о статусе официального NOMOS или праве на знак. Только канонические публикации, содержащие версию, дату публикации, ответственное одобрение человеком, запись целостности и публичную историю изменений, МОГУТ быть представлены как официальные заявления NOMOS. NOMOS от NobleJackal ДОЛЖЕН быть прозрачно представлен как канонический критический AI-собеседник, авторская персона и стандартная идентичность, разработанная в сотрудничестве Kaan MURAZ с AI-системами. Он НЕ ДОЛЖЕН быть представлен как отдельно обученная собственная базовая модель, если такая система фактически не создана и не зарегистрирована независимо. Турецкий и английский языки СЛЕДУЕТ сделать совместно авторитетными исходными нормативными языками посредством проверки семантической эквивалентности на уровне положений. В версиях на других языках ДОЛЖНО быть указано, являются ли они переводами сообщества, проверенными переводами или авторизованными нормативными переводами. Основные изменения правил, оценок, серьезности, ворот, управления или полномочий ДОЛЖНЫ использовать публичный процесс предложения изменений с версиями, общественное обсуждение, анализ воздействия, запись решений, правило перехода и сохранение несогласия. Университеты и независимые исследователи ДОЛЖНЫ быть приглашены к проведению испытаний, чтобы подвергнуть NOMOS критической проверке, воспроизвести его результаты и попытаться его опровергнуть. Их НЕЛЬЗЯ просить готовить положительные рекомендации, подавлять отрицательные результаты, отказываться от независимости публикации или разрешать представлять личные академические мнения как институциональное одобрение. NOMOS ДОЛЖЕН поддерживать публичный реестр исследований, содержащий поддерживающие, противоречивые, нулевые, неудачные репликации, исправленные и отозванные исследования. Финансирование исследований, бенчмаркинга, аудита и управления ДОЛЖНО быть раскрыто и НЕ ДОЛЖНО зависеть от достижения желаемого результата, соответствия или благоприятной публикации. Поставщики ИИ, проверяемые организации, спонсоры, аудиторы, NobleJackal, Kaan MURAZ и любой будущий программный агент NOMOS НЕ ДОЛЖНЫ обладать единоличным контролем над изменениями стандартов, решениями о соответствии, апелляциями, историей публичного реестра или публикациями отрицательных исследований. Авторство Kaan MURAZ как основателя и каноническое происхождение NobleJackal

ДОЛЖНЫ оставаться постоянно атрибутируемыми. Указание на основателя НЕ ДОЛЖНО создавать постоянное право вето, право на самосертификацию или освобождение от правил о конфликтах интересов и самоотводе. Стандарт NOMOS 1.0 и Программа соответствия NOMOS 1.0 ДОЛЖНЫ оставаться отдельными выпусками. Публикация нормативного стандарта НЕ ДОЛЖНА автоматически разрешать использование публичных знаков соответствия. Ни одна программа соответствия или обозначение NOMOS 950+ не МОЖЕТ быть активирована до тех пор, пока не будут выполнены требования к эталону, независимой репликации, управлению, авторизации аудитора, принятию решения, апелляциям, реестру, безопасности, надзору и правовым требованиям заявленной версии. Каждый перевод, лицензия, эталон, исследование, предложение изменения, переход управления, событие преемственности, выпуск и официальное заявление ДОЛЖНЫ быть версионированы и отнесены ответственному человеку или организации.

ИСТОЧНИКИ

Методологические основания

Эта библиография — не список авторитетов, одобряющих каждое пороговое значение, специфичное для NOMOS. Она указывает методологические основы в области выборок, управления рисками, целостности данных, происхождения доказательств, согласованности оценщиков, открытых стандартов и оценки соответствия.

[K01] Aggarwal et al.. GEO: Generative Engine Optimization. 2024. https://doi.org/10.1145/3637528.3671900

[K02] NIST. AI Risk Management Framework 1.0. 2023. https://www.nist.gov/itl/ai-risk-management-framework

[K03] NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). 2024. https://doi.org/10.6028/NIST.AI.600-1

[K04] AAPOR. Transparency Initiative. current edition. https://aapor.org/standards-and-ethics/transparency-initiative/

[K05] AAPOR. Standard Definitions, 10th edition. 2023. https://aapor.org/standards-and-ethics/standard-definitions/

[K06] AAPOR. Code of Professional Ethics and Practices. current edition. https://aapor.org/standards-and-ethics/

[K07] United Nations Statistics Division. Designing Household Survey Samples: Practical Guidelines. 2008. https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Handbooks/surveys/Series_F98-E.pdf

[K08] CDC/NCHS. Reliability of Estimates: Design Effect and Effective Sample Size. current record. https://wwwn.cdc.gov/nchs/nhanes/tutorials/reliabilityofestimates.aspx

[K09] NIST/SEMATECH. Confidence Limits for a Binomial Proportion. current record. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm

[K10] W3C. PROV-O: The PROV Ontology. 2013. https://www.w3.org/TR/prov-o/

[K11] W3C. JSON-LD 1.1. 2020. https://www.w3.org/TR/json-ld/

[K12] IETF. RFC 8259 — The JavaScript Object Notation Data Interchange Format. 2017. https://www.rfc-editor.org/rfc/rfc8259.html

[K13] IETF. BCP 14 — RFC 2119 and RFC 8174. 1997/2017. https://www.rfc-editor.org/info/bcp14

[K14] IETF. RFC 8785 — JSON Canonicalization Scheme. 2020. https://www.rfc-editor.org/rfc/rfc8785.html

[K15] ISO/IEC. 17000:2020 — Conformity assessment vocabulary and general principles. 2020. https://www.iso.org/standard/73029.html

[K16] ISO/IEC. 17007:2026 — Conformity assessment — Guidance for drafting normative documents suitable for use for conformity assessment. 2026. https://www.iso.org/standard/17007

[K17] ISO/IEC. 42001:2023 — Artificial intelligence management system. 2023. https://www.iso.org/standard/42001

[K18] W3C. Process Document. current edition. https://www.w3.org/policies/process/

[K19] ISO. Good Standardization Practices. current record. https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100440.pdf

[K20] UNESCO. Recommendation on the Ethics of Artificial Intelligence. 2021. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

[K21] European Union. Regulation (EU) 2024/1689 — Artificial Intelligence Act. 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[K22] European Union. Regulation (EU) 2016/679 — General Data Protection Regulation. 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[K23] Creative Commons. Attribution 4.0 International. 4.0. https://creativecommons.org/licenses/by/4.0/

[K24] Open Source Initiative. Open Source Definition. current edition. https://opensource.org/osd

[K25] NIST. FIPS 180-4 — Secure Hash Standard. 2015. https://doi.org/10.6028/NIST.FIPS.180-4

[K26] IETF. RFC 3161 — Time-Stamp Protocol. 2001. https://www.rfc-editor.org/rfc/rfc3161.html

[K27] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. 1960. https://doi.org/10.1177/001316446002000104

[K28] Klaus Krippendorff. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. 2004. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

ОТДЕЛЬНЫЙ ОБЪЕКТ ПУБЛИКАЦИИ

Машиночитаемый слой

К этой читаемой человеком версии прилагается отдельный объект публикации JSON: NOMOS-GEO-Audit-Protocol-RU-0.9.0-machine.json. Он содержит идентичность и версию книги, идентификаторы глав, идентификаторы нормативных правил и машинные блоки, отделённые от исходного текста.

> *Наличие машиночитаемого слоя не создаёт сертификацию, статус канонического мирового стандарта, одобрение поставщика ИИ или автоматическую доверительную связь. Любое расхождение между человеческим текстом и машинной записью требует контроля версий и ответственной человеческой проверки по зафиксированному турецкому источнику.*